Adaptive estimation of stationary Gaussian fields

Nicolas Verzelen

The Annals of Statistics
2010, Vol. 38, No. 3, 1363–1402
DOI: 10.1214/09-AOS751

ADAPTIVE ESTIMATION OF STATIONARY GAUSSIAN FIELDS

BY NICOLAS VERZELEN[1]

INRA and SUPAGRO

We study the nonparametric covariance estimation of a stationary Gaussian field X observed on a regular lattice. In the time series setting, some procedures like AIC are proved to achieve optimal model selection among autoregressive models. However, there exists no such equivalent results of adaptivity in a spatial setting. By considering collections of Gaussian Markov random fields (GMRF) as approximation sets for the distribution of X, we introduce a novel model selection procedure for spatial fields. For all neighborhoods m in a given collection $\mathcal{M}$, this procedure first amounts to computing a covariance estimator of X within the GMRFs of neighborhood m. Then it selects a neighborhood $\widehat{m}$ by applying a penalization strategy. The so-defined method satisfies a nonasymptotic oracle-type inequality. If X is a GMRF, the procedure is also minimax adaptive to the sparsity of its neighborhood. More generally, the procedure is adaptive to the rate of approximation of the true distribution by GMRFs with growing neighborhoods.

1. Introduction. In this paper, we study the estimation of the distribution of a stationary Gaussian field $X = (X_{[i,j]})_{(i,j)\in\Lambda}$ indexed by the nodes of a square lattice Λ of size $p \times p$. This problem is often encountered in spatial statistics or in image analysis.

Various estimation methods have been proposed to handle this question. Most of them fall into two categories. On the one hand, one may consider direct covariance estimation. A traditional approach amounts to computing an empirical variogram and then fitting a suitable parametric variogram model such as the exponential or Matérn model (Cressie [10], Chapter 2). Some procedures also apply to nonregular lattices. However, a bad choice of the variogram model may lead to poor results. The issue of variogram model selection has not been completely solved yet although some procedures based on cross-validation have been proposed. See [10], Section 2.6.4, for a discussion. Most of the nonparametric (Hall, Fisher and Hoffmann [19]) and semiparametric (Im, Stein and Zhu [21]) methods are based on the spectral representation of the field. To our knowledge, these procedures have not yet been shown to achieve adaptiveness; that is, their rate of convergence does not adapt to the *complexity* of the correlation functions.

Received January 2009; revised September 2009.

[1]Research mostly carried out at Univ. Paris-Sud (Laboratoire de Matématiques, CNRS-UMR 8628).

AMS 2000 subject classifications. Primary 62H11; secondary 62M40.

Key words and phrases. Gaussian field, Gaussian Markov random field, model selection, pseudo-likelihood, oracle inequalities, minimax rate of estimation.

An alternative approach to the problem amounts to considering the conditional distribution at one node given the remaining nodes. This point of view is closely connected to the notion of *Gaussian Markov random field* (GMRF). Let $\mathcal{G}$ be a graph whose vertex set is Λ. The field X is GMRF with respect to $\mathcal{G}$ if it satisfies the following property: for any node $(i, j) \in \Lambda$, conditionally to the set of variables $X_{[k,l]}$ such that (k, l) is a neighbor of (i, j) in $\mathcal{G}$, $X_{[i,j]}$ is independent from all the remaining variables. GMRFs are also sometimes called Gaussian graphical models. A huge literature develops around this subject since Gaussian graphical models are promising tools to analyze complex high-dimensional systems involved, for instance, in postgenomic data. In other applications, GMRFs are relevant because they allow one to perform a Markov chain Monte Carlo run quickly using Markov properties (e.g., [31]). See Lauritzen [24] or Edwards [14] for introductions to Gaussian graphical models and Markov properties. In the sequel, we assume that the node $(0, 0)$ belongs to Λ. Since we assume that the field X is stationary, defining a graph $\mathcal{G}$ is equivalent to defining the neighborhood m of the node $(0, 0)$. Indeed, the neighborhood of any node $(i, j) \in \Lambda$ is the transposition of m by (i, j). In the sequel, we call m *the neighborhood* of a GMRF. If the neighborhood is empty, then the Markov property states that the components of X are all independent. Alternatively, any zero-mean Gaussian stationary field is a GMRF with respect to the complete neighborhood [i.e., containing all the nodes except $(0, 0)$].

Numerous papers have been devoted to parametric estimation for stationary GMRFs with a known neighborhood. The authors have derived their asymptotic properties of such estimators (see [3, 5, 16]). If the field X is assumed to be a GMRF with respect to a *known* neighborhood, in each of these works, the issue of neighborhood selection has been less studied. Besag and Kooperberg [4], Rue and Tjelmeland [31], Song, Fuentes and Ghosh [33] and Cressie and Verzelen [11] have tackled the problem of *approximating* the distribution of a Gaussian field by a GMRF, but this requires the knowledge of the true distribution. Guyon and Yao have stated in [18] necessary conditions and sufficient conditions for a model selection procedure to choose asymptotically the true neighborhood of a GMRF with probability one.

In this paper, we study a nonparametric estimation procedure based on neighborhood selection. In short, we select a *suitable* neighborhood and estimate the distribution of X in the space of stationary GMRFs with respect to this neighborhood. The objective is not to estimate the "true" neighborhood. We rather want to select a neighborhood that allows to estimate *well* the distribution of X (i.e., to minimize a risk). In fact, we do not even assume that the true correlation of X corresponds to a GMRF. This estimation procedure is relevant for two main reasons:

- To our knowledge, it is the first nonparametric estimator in a spatial setting which achieves adaptive rates of convergence.

- In most of the statistical applications where GMRFs are involved, the neighborhood is a priori unknown. Our procedure allows one to select a "good" neighborhood.

Our problem on a two-dimensional field has a natural one-dimensional counterpart in time series analysis. It is indeed known that an auto-regressive process (AR) of order p is also a GMRF with $2p$ nearest neighbors and reciprocally (see [17], Section 1.3). In this one-dimensional setting, our issue reformulates as follows: how can we select the order of an AR to estimate well the distribution of a time series? It is known that order selection by minimization of criteria like AICC, AIC or FPE satisfy asymptotically oracle inequalities (Shibata [32] and Hurvich and Tsai [20]). We refer to Brockwell and Davis [9] and McQuarrie and Tsai [26] for detailed discussions. However, one cannot readily extend these results to a spatial setting because of computational and theoretical difficulties.

In the rest of this introduction, we further describe the framework and we summarize the main results of the paper.

1.1. *Conditional regression.* Let us now make precise the notation and present the ideas underlying our approach. In the sequel, Λ stands for the toroidal lattice of size $p \times p$. We consider the random field $X = (X_{[i,j]})_{1 \le i,j \le p}$ indexed by the nodes of Λ. Additionally, X^v refers to the vectorialized version of X with the convention $X_{[i,j]} = X^v_{[(i-1)\times p + j]}$ for any $1 \le i, j \le p$. Using this new notation amounts to "forgetting" the spatial structure of X and allows one to get into a more classical statistical framework. For the sake of simplicity, the components of X are defined modulo p in the remainder of the paper.

Throughout this paper, we assume the field X is centered. In practice, the statistician has to first subtract some parametric form of the mean value. Hence, the vector X^v follows a zero-mean Gaussian distribution $\mathcal{N}(0, \Sigma)$ where the $p^2 \times p^2$ matrix Σ is nonsingular but unknown. Also, we suppose that the field X is stationary on the torus Λ. More precisely, for any $r > 0$, any $(i, j) \in \{1, \dots, p\}^2$ and any $(k_1, l_1), \dots, (k_r, l_r) \in \{1, \dots, p\}^{2r}$, it holds that

$$\big(X_{[k_1,l_1]}, \dots, X_{[k_r,l_r]}\big) \sim \big(X_{[k_1+i,l_1+j]}, \dots, X_{[k_r+i,l_r+j]}\big).$$

We observe $n \ge 1$ i.i.d. replications of the vector X^v. In the sequel, $\mathbf{X^v}$ denotes the $p^2 \times n$ matrix of the n observations of X^v. For any $1 \le i \le n$, the $p \times p$ matrix $\mathbf{X}_i$ stands for the ith observation of the field X. All these notations are recalled in Table 1. In practice, the number of observations n often equals one. Our goal is to estimate the matrix Σ.

We sometimes assume that the field X is isotropic. Let G be the group of vector isometries of the unit square. For any node $(i, j) \in \Lambda$ and any isometry $g \in G$, $g.(i, j)$ stands for the image of (i, j) in Λ under the action of g. We say that X is isotropic on Λ if for any $r > 0$, $g \in G$, and $(k_1, l_1), \dots, (k_r, l_r) \in \{1, \dots, p\}^{2r}$,

$$\big(X_{[k_1,l_1]}, \dots, X_{[k_r,l_r]}\big) \sim \big(X_{[g.(k_1,l_1)]}, \dots, X_{[g.(k_r,l_r)]}\big).$$

As mentioned earlier, we aim at estimating the distribution of the field X through a conditional distribution approach. By standard Gaussian derivations (see, for instance, [24], Appendix C), there exists a unique $p \times p$ matrix θ such that $\theta_{[0,0]} = 0$ and

$$X_{[0,0]} = \sum_{(i,j)\in\Lambda\setminus\{(0,0)\}} \theta_{[i,j]} X_{[i,j]} + \varepsilon_{[0,0]}, \tag{1}$$

where the random variable $\varepsilon_{[0,0]}$ follows a zero-mean normal distribution and is independent from the covariates $(X_{[i,j]})_{(i,j)\in\Lambda\setminus\{(0,0)\}}$. Equation (1) describes the conditional distribution of $X_{[0,0]}$ given the remaining variables. Since the field X is stationary, the matrix θ also satisfies $\theta_{[i,j]} = \theta_{[-i,-j]}$ for any $(i, j) \in \Lambda$. Let us note σ^2, the conditional variance of $X_{[0,0]}$, and I_{p^2}, the identity matrix of size p^2. The matrix θ is closely related to the covariance matrix Σ of X^v through the following property:

$$\Sigma = \sigma^2 \big(I_{p^2} - C(\theta)\big)^{-1}, \tag{2}$$

where the $p^2 \times p^2$ matrix $C(\theta)$ is defined as $C(\theta)_{[i_1(p-1)+j_1, i_2(p-1)+j_2]} := \theta_{[i_2-i_1, j_2-j_1]}$ for any $1 \le i_1, i_2, j_1, j_2 \le p$. The matrix $(I_{p^2} - C(\theta))$ is called the partial correlation matrix of the field X. The so-defined matrix $C(\theta)$ is symmetric block circulant with $p \times p$ blocks as stated below. We refer to [29], Section 2.6 or the book of Gray [15] for definitions and main properties on circulant and block circulant matrices.

LEMMA 1.1. *Let θ be a square matrix of size p such that*

$$\text{for any } 1 \le i, j \le p \qquad \theta_{[i,j]} = \theta_{[-i,-j]}; \tag{3}$$

then the matrix $C(\theta)$ is symmetric block circulant with $p \times p$ blocks. Conversely, if B is a $p^2 \times p^2$ symmetric block circulant matrix with $p \times p$ blocks, then there exists a square matrix θ of size p satisfying (3) *and such that $B = C(\theta)$.*

A proof is given in the technical Appendix [36]. In conclusion, estimating the matrix Σ/σ^2 amounts to estimating the matrix $C(\theta)$ which is also equivalent to estimating the $p \times p$ matrix θ. This is why we shall focus on the estimation of the matrix θ.

Let us make precise the set of possible values for θ. In the sequel, Θ denote the vector space of the $p \times p$ matrices that satisfy $\theta_{[0,0]} = 0$ and $\theta_{[i,j]} = \theta_{[-i,-j]}$ for any $(i, j) \in \Lambda$. A matrix $\theta \in \Theta$ corresponds to the distribution of a stationary Gaussian field if and only if the $p^2 \times p^2$ matrix $(I_{p^2} - C(\theta))$ is positive definite. This is why we define the convex subset Θ^+ of Θ by

$$\Theta^+ := \big\{\theta \in \Theta \text{ s.t. } \big(I_{p^2} - C(\theta)\big) \text{ is positive definite}\big\}. \tag{4}$$

The set of covariance matrices of stationary Gaussian fields on Λ with unit conditional variance is therefore in one to one correspondence with the set Θ^+. Let us define the corresponding set Θ^{iso} and $\Theta^{+,\mathrm{iso}}$ for isotropic Gaussian fields:

$$\begin{aligned}\Theta^{\mathrm{iso}} &:= \{\theta \in \Theta, \theta_{[i,j]} = \theta_{[g\cdot(i,j)]}, \forall (i,j) \in \Lambda, \forall g \in G\} \quad \text{and} \\ \Theta^{+,\mathrm{iso}} &:= \Theta^+ \cap \Theta^{\mathrm{iso}}.\end{aligned} \tag{5}$$

1.2. *Model selection*. We have the issue of covariance estimation as an estimation problem for conditional regressions (1). However, the set Θ^+ of admissible parameters for the estimation is huge. The dimension of Θ is indeed of the same order as p^2 whereas we only observe p^2 nonindependent data if n equals one. In order to avoid the curse of dimensionality, it is natural to assume that the target θ is approximately *sparse*.

It is indeed likely that the coefficients $\theta_{[i,j]}$ are *close* to zero for the nodes (i, j) which are *far* from the origin $(0, 0)$. By (1), this means that $X_{[0,0]}$ is *well* predicted by the covariates $X_{[i,j]}$ whose corresponding nodes (i, j) are close to the origin. In other terms, the true covariance is presumably well approximated by a GMRF with a *reasonable* neighborhood. The main difficulty is that we do not know a priori what "reasonable" means. We want to adapt to the *sparsity* of the matrix θ.

In the sequel, m refers to a subset of $\Lambda \setminus \{0, 0\}$. We call it a model. By (1), the property, "X is a GMRF with respect to the neighborhood m," is equivalent to, "the support of θ is included in m." We are given a nested collection $\mathcal{M}$ of models. For any of these models $m \in \mathcal{M}$, we compute $\widehat{\theta}_{m,\rho_1}$, the conditional least squares estimator (CLS), of θ for the model m by maximizing the pseudolikelihood over a subset of matrices θ whose support is included in m. These estimators, as well as their dependency on the quantity ρ_1, are defined in Section 2.

The model m that minimizes the risk of $\widehat{\theta}_{m,\rho_1}$ over the collection $\mathcal{M}$ is called an oracle and is noted m^*. In practice, this model is unknown and we have to estimate it. The art of model selection is to pick a model $m \in \mathcal{M}$ that is large enough to enable a good approximation of θ but is small enough so that the variance of $\widehat{\theta}_{m,\rho_1}$ is small. Let us reformulate the approach in terms of GMRFs: given a collection $\mathcal{M}$ of neighborhoods, we compute an estimator of θ in the set of GMRFs with neighborhood m for any $m \in \mathcal{M}$. Our purpose is to select a suitable neighborhood $\widehat{m}$ so that the estimator $\widehat{\theta}_{\widehat{m}}$ has a risk as small as possible.

A classical method to estimate a *good model* $\widehat{m}$ is achieved through *penalization* with respect to the size of the models. In the following expression, $\gamma_{n,p}(\cdot)$ stands for the CLS empirical contrast that we shall define in Section 2. We select a model $\widehat{m}$ by minimizing the criterion,

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} [\gamma_{n,p}(\widehat{\theta}_{m,\rho_1}) + \mathrm{pen}(m)], \tag{6}$$

where $\mathrm{pen}(\cdot)$ denotes a positive function defined on $\mathcal{M}$. In this paper, we prove that under a suitable choice of the penalty function $\mathrm{pen}(\cdot)$, the risk of the estimator $\widehat{\theta}_{\widehat{m}}$ is as small as possible.

1.3. *Risk bounds and adaptation.* We shall assess our procedure using two different loss functions. First, we introduce the loss function $l(\cdot,\cdot)$ that measures how well we estimate the conditional distribution (1) of the field. For any $\theta_1, \theta_2 \in \Theta$, the distance $l(\theta_1, \theta_2)$ is defined by

$$l(\theta_1,\theta_2) := \frac{1}{p^2}\operatorname{tr}\big[\big(C(\theta_1)-C(\theta_2)\big)\Sigma\big(C(\theta_1)-C(\theta_2)\big)\big]. \tag{7}$$

Let us reformulate $l(\theta_1, \theta_2)$ in terms of conditional expectation,

$$l(\theta_1,\theta_2) = \mathbb{E}_\theta\big\{\big[\mathbb{E}_{\theta_1}\big(X_{[0,0]}|X_{\Lambda\setminus\{0,0\}}\big) - \mathbb{E}_{\theta_2}\big(X_{[0,0]}|X_{\Lambda\setminus\{0,0\}}\big)\big]^2\big\},$$

where $\mathbb{E}_\theta(\cdot)$ stands for the expectation with respect to the distribution of X^v, $\mathcal{N}(0, \sigma^2(I_{p^2} - C(\theta))^{-1})$. Hence, $l(\widehat{\theta}, \theta)$ corresponds the mean squared prediction loss which is often used in the random design regression framework, in time series analysis [20] or in spatial statistics [33]. Moreover, the loss function $l(\widehat{\theta}, \theta)$ is also connected to the notion of kriging error. The kriging predictor (Stein [34]) of $X_{[0,0]}$ is defined as the best linear combination of the covariates $(X_{[k,l]})_{(k,l)\in\Lambda\setminus\{(0,0\}}$ for predicting the value $X_{[0,0]}$. By (1), this predictor is exactly $\sum_{(k,l)\in\Lambda\setminus\{(0,0\}}\theta_{[k,l]}X_{[k,l]}$, and the mean squared prediction error is σ^2. If we do not know θ but we are given an estimator $\widehat{\theta}$, then the corresponding kriging predictor $\sum_{(k,l)\in\Lambda\setminus\{(0,0\}}\widehat{\theta}_{[k,l]}X_{[k,l]}$ has a mean squared prediction error equal to $\sigma^2 + l(\widehat{\theta}, \theta)$. Kriging is a key concept in spatial statistics, and it is therefore interesting to consider a loss function that measures the kriging performances when one estimates θ.

We shall also assess our results using the Frobenius distance noted $\|\cdot\|_F$ and defined by $\|A\|_F^2 := \sum_{1\le i,j\le p} A_{[i,j]}^2$. Observe that the Frobenius distance $\|\theta_1 - \theta_2\|_F^2$ also equals the Frobenius distance between the partial correlation matrices $(I_{p^2} - C(\theta_1))$ and $(I_{p^2} - C(\theta_2))$ (up to a factor p^2)

$$\|\theta_1-\theta_2\|_F^2 = \frac{1}{p^2}\big\|\big(I_{p^2}-C(\theta_1)\big)-\big(I_{p^2}-C(\theta_2)\big)\big\|_F^2. \tag{8}$$

Our aim is then to define a suitable penalty function $\operatorname{pen}(\cdot)$ in (6) so that the estimator $\widehat{\theta}_{\widehat{m},\rho_1}$ performs almost as well as the oracle estimator $\widehat{\theta}_{m^*,\rho_1}$. For any model $m \in \mathcal{M}$, we define θ_{m,ρ_1} as the matrix which minimizes the loss $l(\theta', \theta)$ over the sets of matrices θ' corresponding to model m. The loss $l(\theta_{m,\rho_1}, \theta)$ is called the *bias*. Our main result is stated in Section 3. We provide a condition on the penalty function $\operatorname{pen}(\cdot)$, so that the selected estimator satisfies a risk bound of the form

$$\mathbb{E}_\theta[l(\widehat{\theta}_{\widehat{m},\rho_1},\theta)] \le L\inf_{m\in\mathcal{M}}\left[l(\theta_{m,\rho_1},\theta)+\varphi_{\max}(\Sigma)\frac{\operatorname{Card}(m)}{np^2}\right], \tag{9}$$

where $\varphi_{\max}(\Sigma)$ is the largest eigenvalue of Σ, and $\operatorname{Card}(\cdot)$ stands for the cardinality. Contrary to most results in a spatial setting, this upper bound on the risk is nonasymptotic and holds in a general setting. The term $\varphi_{\max}(\Sigma)\operatorname{Card}(m)/(np^2)$

grows linearly with the size of m and goes to 0 with n and p. In Section 4, we prove that the variance term of a model m is of the same order as $\varphi_{\max}(\Sigma)\operatorname{Card}(m)/(np^2)$. Hence, the bound (9) tells us that the risk of $\widehat{\theta}_{\widehat{m},\rho_1}$ is smaller than a quantity which is the same order as the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m^*,\rho_1},\theta)]$ of the oracle m^*. We say that the selected estimator achieves an *oracle-type inequality*.

In Section 4, we bound the asymptotic expectations $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1},\theta)]$ and connect them to the variance terms in bound (9). As a consequence, we prove that under mild assumptions on the target θ, the upper bound (9) is optimal from the asymptotic point of view (up to a multiplicative numerical constant). We discuss the assumptions in Section 5. In Section 6, we compute nonasymptotic minimax lower bounds with respect to the loss functions $l(\cdot,\cdot)$ and $\|\cdot\|_F^2$. We then derive that under mild assumptions, our estimator $\widehat{\theta}_{\widehat{m},\rho_1}$ is minimax adaptive to the sparsity of θ and minimax adaptive to the decay of θ.

To our knowledge, these are the first oracle-type inequalities in a spatial setting. The computation of the minimax rates of convergence is also new. Moreover, most of our results are nonasymptotic. Although we have considered a square on the two-dimensional lattice, our method straightforwardly extends to any d-dimensional toroidal rectangle with $d \geq 1$. In the one-dimensional setting, we retrieve a oracle-type inequality that is close to the work of Shibata [32]. Yet, he has stated an asymptotic oracle inequality for the estimation of autoregressive processes. In contrast, our result applies on a torus and is only optimal up to constants but it is nonasympotic, and, most of all, it applies for higher-dimensional lattices. In Section 7, we further discuss the advantages and the weak points of our method. Moreover, we mention the extensions and the simulations made in a subsequent paper [37]. All the proofs are postponed to Section 8 and to the Appendix [36].

1.4. *Some notation.* Throughout this paper, $L, L_1, L_2, \ldots$ denote constants that may vary from line to line. The notation $L(\cdot)$ specifies the dependency on some quantities. For any matrix A, $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$, respectively, refer the largest eigenvalue and the smallest eigenvalues of A. We recall that $\|A\|_F$ is the Frobenius norm of A. For any matrix θ of size p, $\|\theta\|_1$ stands for the sum of of the absolute values of the components of θ; we call it its l_1 norm. In the sequel, 0_p is the square matrix of size p whose indices are 0. Given $\rho > 0$, the ball $\mathcal{B}_1(0_p;\rho)$ is defined as the set of square matrices of size p whose l_1 norm is smaller than ρ. Finally, Table 1 gathers the notation involving X.

TABLE 1
Notations for the random field and the data

X	Matrix of size $p \times p$	Random field
X^v	Vector of length p^2	Vectorialized version of X
$\mathbf{X^v}$	Matrix of size $p^2 \times n$	Observations of X^v
$\mathbf{X}_i$	Matrix of size $p \times p$	ith observation of the field X

2. Model selection procedure. In this section, we formally define our model selection procedure.

2.1. *Collection of models.* For any node (i, j) belonging to the lattice Λ, let us define the toroidal norm by

$$|(i, j)|_t^2 := [i \wedge (p - i)]^2 + [j \wedge (p - j)]^2.$$

We aim at selecting a "good" neighborhood for the GMRF. Since X corresponds to some "spatial" process, it is natural to assume that nodes that are close to $(0, 0)$ are more likely to be significant. This is why we restrict ourselves in the sequel to the collection $\mathcal{M}_1$ of neighborhoods.

DEFINITION 2.1. A subset $m \subset \Lambda \setminus \{(0, 0)\}$ belongs to $\mathcal{M}_1$ if there exists a number $r_m > 1$ such that

$$m = \big\{(i, j) \in \Lambda \setminus \{(0, 0)\} \text{ s.t. } |(i, j)|_t \leq r_m\big\}. \tag{10}$$

The collection $\mathcal{M}_1$ is totally ordered with respect to the inclusion and we therefore order our models $m_0 \subset m_1 \subset \cdots \subset m_i \cdots$. For instance, m_0 corresponds to the empty neighborhood whereas m_1 stands for the neighborhood of size 4. See Figure 1 for other examples.

For any model $m \in \mathcal{M}_1$, we define the vector space Θ_m as the subset of the elements of Θ whose support is included in m. We recall that Θ is defined in Section 1.1. Similarly Θ_m^{iso} is the subset of Θ^{iso} whose support is included in m. The dimensions of Θ_m and Θ_m^{iso} are, respectively, noted d_m and d_m^{iso}. Since we

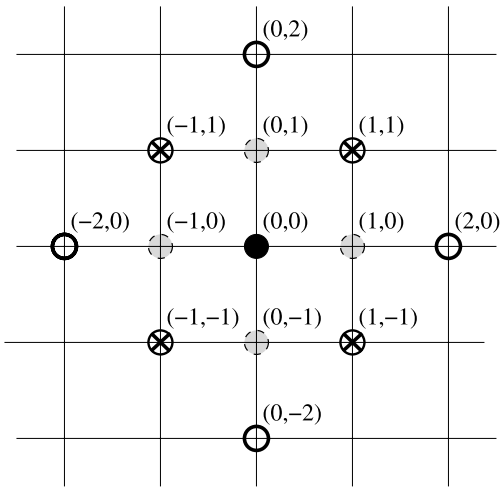

FIG. 1. *Examples of models. The four gray nodes refer to* m_1. *The model* m_2 *also contains the nodes with a cross whereas* m_3 *contains all the nodes except* $(0, 0)$.

aim at estimating the positive matrix $(I_{p^2} - C(\theta))$, we shall consider the convex subsets of Θ_m^+ and $\Theta_m^{+,\mathrm{iso}}$ that correspond to nonnegative precision matrices,

$$\Theta_m^+ := \Theta_m \cap \Theta^+ \quad \text{and} \quad \Theta_m^{+,\mathrm{iso}} := \Theta_m^{\mathrm{iso}} \cap \Theta^{+,\mathrm{iso}}. \tag{11}$$

For instance, the set $\Theta_{m_1}^+$ is in one-to-one correspondence with the sets of GMRFs whose neighborhood is made of the four nearest neighbors. Similarly, $\Theta_{m_1}^+$ is in one-to-one correspondence with the GMRFs with eight nearest neighbors. In our estimation procedure, we shall restrict ourselves to precision matrices whose largest eigenvalue is upper bounded by a constant. This is why we define the subsets Θ_{m_2,ρ_1}^+ and $\Theta_{m,\rho_1}^{+,\mathrm{iso}}$ for any $\rho_1 \geq 2$:

$$\Theta_{m,\rho_1}^+ := \{\theta \in \Theta_m^+, \varphi_{\max}(I_{p^2} - C(\theta)) < \rho_1\}, \tag{12}$$

$$\Theta_{m,\rho_1}^{+,\mathrm{iso}} := \{\theta \in \Theta_m^{+,\mathrm{iso}}, \varphi_{\max}(I_{p^2} - C(\theta)) < \rho_1\}. \tag{13}$$

Finally, we need a generating family of the spaces Θ_m and Θ_m^{iso}. For any node $(i, j) \in \Lambda \setminus \{(0, 0)\}$, let us define the $p \times p$ matrix $\Psi_{i,j}$ as

$$\Psi_{i,j[k,l]} := \begin{cases} 1, & \text{if } (k, l) = (i, j) \text{ or } (k, l) = -(i, j), \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Hence, Θ_m is generated by the matrices $\Psi_{i,j}$ for which (i, j) belongs to m. Similarly, for any $(i, j) \in \Lambda \setminus \{(0, 0)\}$, let us define the matrix $\Psi_{i,j}^{\mathrm{iso}}$ by

$$\Psi_{i,j[k,l]}^{\mathrm{iso}} := \begin{cases} 1, & \text{if } \exists g \in G, (k, l) = g.(i, j), \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

2.2. *Estimation by conditional least squares* (*CLS*). Let us turn to the conditional least squares estimator. For any $\theta' \in \Theta^+$, the criterion $\gamma_{n,p}(\theta')$ is defined by

$$\gamma_{n,p}(\theta') := \frac{1}{np^2} \sum_{i=1}^{n} \sum_{1 \leq j_1, j_2 \leq p} \Bigg(\mathbf{X}_{i[j_1,j_2]} - \sum_{(l_1,l_2) \in \Lambda \setminus \{(0,0)\}} \theta'_{[l_1,l_2]} \mathbf{X}_{i[j_1+l_1, j_2+l_2]} \Bigg)^2. \tag{16}$$

In a nutshell, $\gamma_{n,p}(\theta')$ is a least squares criterion that allows one to perform the simultaneous linear regression of all $\mathbf{X}_{i[j_1,j_2]}$ with respect to the covariates $(\mathbf{X}_{i[l_1,l_2]})_{(l_1,l_2) \neq (j_1,j_2)}$. The advantage of this criterion is that it does not require the computation of a determinant of a huge matrix as for the likelihood. We shall often use an alternative expression of $\gamma_{n,p}(\theta')$ in terms of the factor $C(\theta')$ and the empirical covariance matrix $\overline{\mathbf{X}^{\mathbf{v}}\mathbf{X}^{\mathbf{v}*}}$,

$$\gamma_{n,p}(\theta') = \frac{1}{p^2} \operatorname{tr}\big[(I_{p^2} - C(\theta'))\overline{\mathbf{X}^{\mathbf{v}}\mathbf{X}^{\mathbf{v}*}}(I_{p^2} - C(\theta'))\big]. \tag{17}$$

One proves the equivalence between these two expressions by coming back to the definition of $C(\theta')$. Let $\rho_1 > 2$ be fixed. For any model $m \in \mathcal{M}$, we compute the CLS estimators $\widehat{\theta}_{m,\rho_1}$ and $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ by minimizing the criterion $\gamma_{n,p}(\cdot)$ as follows:

$$\widehat{\theta}_{m,\rho_1} := \operatorname*{arg\,min}_{\theta' \in \overline{\Theta^+_{m,\rho_1}}} \gamma_{n,p}(\theta') \quad \text{and} \quad \widehat{\theta}^{\mathrm{iso}}_{m,\rho_1} := \operatorname*{arg\,min}_{\theta' \in \overline{\Theta^{+,\mathrm{iso}}_{m,\rho_1}}} \gamma_{n,p}(\theta'), \tag{18}$$

where $\overline{A}$ stands for the closure of the set A. The existence and the uniqueness of $\widehat{\theta}_{m,\rho_1}$ and $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ are ensured by the following lemma.

LEMMA 2.2. *For any $\theta \in \Theta^+$, $\gamma_{n,p}(\cdot)$ is almost surely strictly convex on $\overline{\Theta^+}$.*

The proof is postponed to the Appendix [36]. We discuss the dependency of $\widehat{\theta}_{m,\rho_1}$ on the parameter ρ_1 in Section 5. For stationary Gaussian fields, minimizing the CLS criterion $\gamma_{n,p}(\cdot)$ over a set Θ^+_{m,ρ_1} is equivalent to minimizing the product of the conditional likelihoods $(X_{[i,j]}|X_{-\{i,j\}})$, called *conditional pseudo-likelihood* (CPL),

$$\begin{aligned} p\mathcal{L}_n(\theta', \mathbf{X}^{\mathbf{v}}) &:= \prod_{\substack{1 \le i \le n, \\ (j_1,j_2) \in \Lambda}} \mathcal{L}_{n,\theta'}\big(\mathbf{X}_{i[j_1,j_2]} | (\mathbf{X}_i)_{-\{j_1,j_2\}}\big) \\ &= \big(\sqrt{2\pi}\sigma\big)^{-np^2} \exp\bigg(-\frac{1}{2}\frac{np^2\gamma_{n,p}(\theta')}{\sigma^2}\bigg), \end{aligned}$$

where we recall that σ^2 refers to the conditional variance of any $X_{[i,j]}$. In fact, CLS estimators were first introduced by Besag [2] who call them pseudolikelihood estimators since they minimize the CPL.

Let us define the function $\gamma(\cdot)$ as an infinite sampled version of the CLS criterion $\gamma_{n,p}(\cdot)$,

$$\gamma(\theta') := \mathbb{E}_\theta[\gamma_{n,p}(\theta')] = \mathbb{E}_\theta\bigg[\bigg(X_{[0,0]} - \sum_{(i,j)\neq(0,0)} \theta'_{[i,j]} X_{[i,j]}\bigg)^2\bigg] \tag{19}$$

for any $\theta', \theta \in \Theta^+$. The function $\gamma(\theta')$ measures the prediction error of $X_{[0,0]}$ if one uses $\sum_{(i,j)\neq(0,0)} \theta'_{[i,j]} X_{[i,j]}$ as a predictor. Moreover, it is a special case of the CMLS criterion introduced by Cressie and Verzelen (see [11], (10)) to approximate a Gaussian field by a GMRF. Hence, one may interpret the CLS criterion as a finite sampled version of their approximation method. Observe that the function $\gamma(\cdot)$ is minimized over Θ^+ at the point θ and that $\gamma(\theta) = \mathrm{Var}_\theta(X_{[0,0]}|X_{-\{0,0\}}) = \sigma^2$. Moreover, the difference $\gamma(\theta') - \gamma(\theta)$ equals the loss $l(\theta', \theta)$ defined by (7).

For any model $m \in \mathcal{M}$, we introduce the projections θ_{m,ρ_1} and $\theta^{\mathrm{iso}}_{m,\rho_1}$ as the best approximation of θ in $\overline{\Theta^+_{m,\rho_1}}$ and $\overline{\Theta^{+,\mathrm{iso}}_{m,\rho_1}}$:

$$\theta_{m,\rho_1} := \operatorname*{arg\,min}_{\theta' \in \overline{\Theta^+_{m,\rho_1}}} l(\theta', \theta) \quad \text{and} \quad \theta^{\mathrm{iso}}_{m,\rho_1} := \operatorname*{arg\,min}_{\theta' \in \overline{\Theta^{+,\mathrm{iso}}_{m,\rho_1}}} l(\theta', \theta). \tag{20}$$

Since $\gamma(\cdot)$ is strictly convex on Θ^+, the matrices θ_{m,ρ_1} and $\theta^{\mathrm{iso}}_{m,\rho_1}$ are uniquely defined. By its definition (7), one may interpret $l(\cdot,\cdot)$ as an inner product on the space Θ; therefore, the orthogonal projection of θ onto the convex closed set $\overline{\Theta^+_{m,\rho_1}}$ (resp., $\overline{\Theta^{+,\mathrm{iso}}_{m,\rho_1}}$) with respect to $l(\cdot,\cdot)$ is θ_{m,ρ_1} (resp., $\theta^{\mathrm{iso}}_{m,\rho_1}$). It then follows from a property of orthogonal projections that the loss of $\widehat{\theta}_{m,\rho_1}$ is upper bounded by

$$l(\widehat{\theta}_{m,\rho_1},\theta) \le l(\theta_{m,\rho_1},\theta) + l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}). \tag{21}$$

The first term $l(\theta_{m,\rho_1},\theta)$ accounts for the bias whereas the second term $l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})$ is a variance term. Observe that $\theta \in \Theta^+_m$ does not necessarily imply that the bias $l(\theta_{m,\rho_1},\theta)$ is null because in general $\overline{\Theta^+_m} \neq \overline{\Theta^+_{m,\rho_1}}$. This will be the case only if θ satisfies the following hypothesis:

$$(\mathbb{H}_1): \quad \varphi_{\max}\bigl(I_{p^2} - C(\theta)\bigr) < \rho_1. \tag{22}$$

Assumption $(\mathbb{H}_1)$ is necessary to ensure the existence of a model $m \in \mathcal{M}$ such that the bias is zero (i.e., $\theta_{m,\rho_1} = \theta$). By identity (2), one observes that $(\mathbb{H}_1)$ is equivalent to a lower bound on the smallest eigenvalue of Σ, i.e., $\varphi_{\min}(\Sigma) \le \sigma^2/\rho_1$. We further discuss $(\mathbb{H}_1)$ in Section 5.

For the sake of completeness, we recall the penalization criterion introduced in (6). Given a subcollection of models $\mathcal{M} \subset \mathcal{M}_1$ and a positive function pen: $\mathcal{M} \to \mathbb{R}^+$ that we call a penalty, we select a model as follows:

$$\widehat{m} := \arg\min_{m\in\mathcal{M}}[\gamma_{n,p}(\widehat{\theta}_{m,\rho_1})] + \mathrm{pen}(m)$$

and

$$\widehat{m}^{\mathrm{iso}} := \arg\min_{m\in\mathcal{M}}[\gamma_{n,p}(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1})] + \mathrm{pen}(m).$$

Observe that $\widehat{m}$ and $\widehat{m}^{\mathrm{iso}}$ depend on ρ_1. For the sake clarity, we do not emphasize this dependency in the notation. In the sequel, we write $\widetilde{\theta}_{\rho_1}$ and $\widetilde{\theta}^{\mathrm{iso}}_{\rho_1}$ for $\widehat{\theta}_{\widehat{m},\rho_1}$ and $\widehat{\theta}^{\mathrm{iso},\rho_1}_{\widehat{m}^{\mathrm{iso}}}$.

3. Main result. We now provide a nonasymptotic upper bound for the risk of the estimators $\widetilde{\theta}_{\rho_1}$ and $\widetilde{\theta}^{\mathrm{iso}}_{\rho_1}$. Let us recall that Σ stands for the covariance matrix of X^v.

THEOREM 3.1. *Let K be a positive number larger than a universal constant K_0 and let $\mathcal{M}$ be a subcollection of $\mathcal{M}_1$. If for every model $m \in \mathcal{M}$,*

$$\mathrm{pen}(m) \ge K\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}, \tag{23}$$

then for any $\theta \in \Theta^+$, the estimator $\widetilde{\theta}_{\rho_1}$ satisfies

$$\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1},\theta)] \le L_1(K)\inf_{m\in\mathcal{M}}[l(\theta_{m,\rho_1},\theta) + \mathrm{pen}(m)] + L_2(K)\frac{\rho_1^2\varphi_{\max}(\Sigma)}{np^2}. \tag{24}$$

A similar bound holds if one replaces $\widetilde{\theta}_{\rho_1}$ *by* $\widetilde{\theta}^{\mathrm{iso}}_{\rho_1}$, Θ^+ *by* $\Theta^{+,\mathrm{iso}}$, θ_{m,ρ_1} *by* θ^{iso}_m, *and* d_m *by* d^{iso}_m.

The proof is postponed to Section 8.2. It is based on a novel concentration inequality for suprema of Gaussian chaos stated in Section 8.1. The constant K_0 is made explicit in the proof. Observe that the theorem holds for any n, any p and that we have not performed any assumption on the target $\theta \in \Theta^+$ (resp., $\Theta^{+,\mathrm{iso}}$). If the collection $\mathcal{M}$ does not contain the empty model, one gets the more readable upper bound,

$$\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] \le L(K) \inf_{m \in \mathcal{M}} [l(\theta_{m,\rho_1}, \theta) + \mathrm{pen}(m)].$$

This theorem tells us that $\widetilde{\theta}_{\rho_1}$ essentially performs as well as the best trade-off between the bias term $l(\theta_{m,\rho_1}, \theta)$ and $\rho_1^2 \varphi_{\max}(\Sigma) \frac{d_m}{np^2}$ that plays the role of a variance. Here are some additional comments.

REMARK 1. Consider the special case where the target θ belongs to some parametric set Θ^+_m with $m \in \mathcal{M}$. Suppose that the hypothesis $(\mathbb{H}_1)$ defined in (22) is fulfilled. Choosing a penalty $\mathrm{pen}(m) = K\rho_1^2 \varphi_{\max}(\Sigma) \frac{d_m}{np^2}$, we get

$$\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] \le L(K) \rho_1^2 \varphi_{\max}(\Sigma) \frac{d_m}{np^2}. \tag{25}$$

We shall prove in Sections 4.2 and 6.1 that this rate is optimal both from an asymptotic oracle and a minimax point of view. We have mentioned in Section 2.2 that $(\mathbb{H}_1)$ is necessary for bound (25) to hold. If ρ_1 is chosen large enough, then assumption $(\mathbb{H}_1)$ is fulfilled. We do not have access to this minimal ρ_1 that ensures $(\mathbb{H}_1)$, since it requires the knowledge of θ. Nevertheless, we argue in Section 5 that "moderate" values for ρ_1 ensure assumption $(\mathbb{H}_1)$ when the model m is small.

REMARK 2. We have mentioned in the Introduction that our objective was to obtain oracle inequalities of the form

$$\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] \le L(K) \inf_{m \in \mathcal{M}} \mathbb{E}[l(\widehat{\theta}_{m,\rho_1}, \theta)] = L(K) \mathbb{E}[(\widehat{\theta}_{m^*,\rho_1}, \theta)].$$

This is why we want to compare the sum $l(\theta_{m,\rho_1}, \theta) + \mathrm{pen}(m)$ with $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1}, \theta)]$. First, we provide in Section 4.1 a sufficient condition so that the risk $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ decomposes exactly as the sum $l(\theta_{m,\rho_1}, \theta) + \mathbb{E}[l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]$. Moreover, we compute in Section 4.2 the asymptotic variance term $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]$ and compare it with the penalty term $\rho_1^2 \varphi_{\max}(\Sigma) \frac{d_m}{np^2}$. We shall then derive oracle-type inequalities and discuss the dependency of the different bounds on $\varphi_{\max}(\Sigma)$.

REMARK 3. Condition (23) gives a lower bound on the penalty function pen$(\cdot)$ so that the result holds. Choosing a proper penalty term according to (23) therefore requires an upper bound on the largest eigenvalue of Σ. However, such a bound is seldom known in practice. We shall mention in Section 7 a practical method to calibrate the penalty.

A bound similar to (24) holds for the Frobenius distance between the partial correlation matrices $(I_{p^2} - C(\theta))$ and $(I_{p^2} - C(\widetilde{\theta}_{\rho_1}))$.

COROLLARY 3.2. *Assume the same as in Theorem 3.1, except that there is equality in (23). Then*

$$\begin{aligned}
&\mathbb{E}_\theta[\|C(\widetilde{\theta}_{\rho_1}) - C(\theta)\|_F^2] \\
&\quad \le L_1(K)\frac{\varphi_{\max}(\Sigma)}{\varphi_{\min}(\Sigma)} \inf_{m\in\mathcal{M}}\left[\|C(\theta_{m,\rho_1}) - C(\theta)\|_F^2 + \frac{K\rho_1^2 d_m}{n}\right] \\
&\qquad + L_2(K)\frac{\varphi_{\max}(\Sigma)}{\varphi_{\min}(\Sigma)}\frac{\rho_1^2}{n}.
\end{aligned} \tag{26}$$

A similar result holds for isotropic GMRFs.

PROOF. This is a consequence of Theorem 3.1. By definition (7) of the loss function $l(\cdot,\cdot)$, the two following bounds hold:

$$\begin{aligned}
p^2 l(\theta_1,\theta_2) &\ge \varphi_{\min}(\Sigma)\|C(\theta_1) - C(\theta_2)\|_F^2, \\
p^2 l(\theta_1,\theta_2) &\le \varphi_{\max}(\Sigma)\|C(\theta_1) - C(\theta_2)\|_F^2.
\end{aligned}$$

Gathering these bounds with (24) yields the result. □

The same comments as for Theorem 3.1 hold. We may express this Corollary 3.2 in terms of the risk $\mathbb{E}(\|\widetilde{\theta}_{\rho_1} - \theta\|_F^2)$, since $\|C(\theta_1) - C(\theta_2)\|_F^2 = p^2\|\theta_1 - \theta_2\|_F^2$:

$$\begin{aligned}
\mathbb{E}_\theta[\|\widetilde{\theta}_{\rho_1} - \theta\|_F^2] &\le L_1(K)\frac{\varphi_{\max}(\Sigma)}{\varphi_{\min}(\Sigma)} \inf_{m\in\mathcal{M}}\left[\|\theta_{m,\rho_1} - \theta\|_F^2 + \frac{K\rho_1^2 d_m}{np^2}\right] \\
&\quad + L_2(K)\frac{\varphi_{\max}(\Sigma)}{\varphi_{\min}(\Sigma)}\frac{\rho_1^2}{np^2}.
\end{aligned}$$

4. Parametric risk and asymptotic oracle inequalities. In this section, we study the risk of the parametric estimators $\widehat{\theta}_{m,\rho_1}$ in order to assess the optimality of Theorem 3.1.

4.1. *Bias-variance decomposition.* The properties of the parametric estimator $\widehat{\theta}_{m,\rho_1}$ and of the projection θ_{m,ρ_1} differ slightly whether θ_{m,ρ_1} belongs to the open set Θ^+_{m,ρ_1} or to its border. Observe that Hypothesis $(\mathbb{H}_1)$ defined in (22) does not

necessarily imply that projection θ_{m,ρ_1} belongs to Θ_m^+. This is why we introduce condition $(\mathbb{H}_2)$:

$$\theta \in \mathcal{B}_1(0_p, 1) \quad \Longleftrightarrow \quad \|\theta\|_1 < 1. \tag{27}$$

The condition $\|\theta\|_1 < 1$ is equivalent to $(I_{p^2} - C(\theta))$ is strictly diagonally dominant. Condition $(\mathbb{H}_2)$ implies that the largest eigenvalue of $(I_{p^2} - C(\theta))$ is smaller than 2 and therefore that $(\mathbb{H}_1)$ is fulfilled since ρ_1 is supposed larger than 2. We further discuss this assumption in Section 5.

LEMMA 4.1. *Let $\theta \in \Theta^+$ such that $(\mathbb{H}_2)$ holds and let $m \in \mathcal{M}_1$. Then, the minimum of $\gamma(\cdot)$ over Θ_m is achieved in $\Theta_{m,2}^+$. This implies that*

$$\theta_{m,\rho_1} = \operatorname*{arg\,min}_{\theta' \in \Theta_m} \gamma(\theta') \quad \text{and} \quad \gamma(\theta_{m,\rho_1}) = \operatorname{Var}_\theta\big(X_{[0,0]} | X_m\big).$$

Additionally, $\|\theta_{m,\rho_1}\|_1 \leq \|\theta\|_1$. The same results holds for $\theta_{m,\rho_1}^{\mathrm{iso}}$ if θ in $\Theta^{+,\mathrm{iso}}$.

The proof is given in the technical Appendix [36]. The purpose of this property is threefold. First, we derive that assumption $(\mathbb{H}_2)$ ensures that θ_{m,ρ_1} belongs Θ_{m,ρ_1}^+ and that the smallest eigenvalue of $(I_{p^2} - C(\theta_{m,\rho_1}))$ is larger than $1 - \|\theta\|_1$. Second, it allows to express the projection θ_{m,ρ_1} in terms of conditional expectation (Corollary 4.2). Finally, we deduce a bias-variance decomposition of the estimator $\widehat{\theta}_{m,\rho_1}$ (Corollary 4.3). In other words, the equality holds in (21).

COROLLARY 4.2. *Let $\theta \in \Theta^+$ such that $(\mathbb{H}_2)$ holds and let $m \in \mathcal{M}_1$. The projection θ_{m,ρ_1} is uniquely defined by the equation*

$$\mathbb{E}_\theta\big(X_{[0,0]} | X_m\big) = \sum_{(i,j)\in m} \theta_{m,\rho_1[i,j]} X_{[i,j]}$$

and $\theta_{m,\rho_1[i,j]} = 0$ for any $(i, j) \notin m$. Similarly, if $\theta \in \Theta^{+,\mathrm{iso}}$ satisfies $(\mathbb{H}_2)$, then $\theta_{m,\rho_1}^{\mathrm{iso}}$ is uniquely defined by the equation

$$\mathbb{E}_\theta\big(X_{[0,0]} | X_m\big) = \sum_{(i,j)\in m} \theta_{m,\rho_1[i,j]}^{\mathrm{iso}} X_{[i,j]}$$

and $\theta_{m,\rho_1[i,j]}^{\mathrm{iso}} = 0$ for any $(i, j) \notin m$.

Consequently, $\sum_{1\leq i,j\leq p} \theta_{m,\rho_1[i,j]} X_{[i,j]}$ is the best linear predictor of $X_{[0,0]}$ given the covariates $X_{[i,j]}$ with $(i, j) \in m$. This is precisely the definition of the kriging parameters (Stein [34]). Hence, the matrix θ_{m,ρ_1} corresponds to the kriging parameters of $X_{[0,0]}$ with kriging neighborhood's range of r_m. The distance r_m is introduced in Definition 2.1 and stands for the radius of m.

COROLLARY 4.3. *Let $\theta \in \Theta^+$ such that $(\mathbb{H}_2)$ holds and let $m \in \mathcal{M}_1$. The loss of $\widehat{\theta}_{m,\rho_1}$ decomposes as $l(\widehat{\theta}_{m,\rho_1}, \theta) = l(\theta_{m,\rho_1}, \theta) + l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})$. If θ belongs to $\Theta_m^{+,\mathrm{iso}}$ and $(\mathbb{H}_2)$ holds, then we also have the decomposition $l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}, \theta) = l(\theta^{\mathrm{iso}}_{m,\rho_1}, \theta) + l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}, \theta_{m,\rho_1})$.*

A proof is provided in the technical Appendix [36]. If θ does not satisfy assumption $(\mathbb{H}_2)$, then θ_{m,ρ_1} does not necessarily belong to Θ^+_{m,ρ_1}, and there may not be such a bias variance decomposition.

4.2. *Asymptotic risk.* In this section, we evaluate the risk of each estimator $\widehat{\theta}_{m,\rho_1}$ and use it as a benchmark to assess the result of Theorem 3.1. We have mentioned in Corollary 4.3 that under $(\mathbb{H}_2)$ the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ decomposes into the sum of the bias $l(\theta_{m,\rho_1}, \theta)$ and a variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]$. If this last quantity is of the same order as the penalty $\mathrm{pen}(m)$ introduced in (23), then Theorem 3.1 yields an oracle inequality. However, we are unable to express this variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]$ in a simple form. This is why we restrict ourselves to study the risks when n tends to infinity. Nevertheless, these results give us some hints to appreciate the strength and the weaknesses of Theorem 3.1 and the upper bound (25).

In the following proposition, we adapt a result of Guyon [17], Section 4.3.2 to obtain an asymptotic expression of the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]$. We first need to introduce some new notation. For any model m in the collection $\mathcal{M}_1 \setminus \{\varnothing\}$, we fix a sequence $(i_k, j_k)_{k=1,\dots,d_m}$ of integers such that $(\Psi_{i_1,j_1}, \dots, \Psi_{i_{d_m},j_{d_m}})$ is a basis of the space Θ_m. Then $\chi_{m[0,0]}$ stands for the random vector of size d_m that contains the neighbors of $X_{[0,0]}$

$$\chi^*_{m[0,0]} := [\mathrm{tr}(\Psi_{i_1,j_1} X^\upsilon), \dots, \mathrm{tr}(\Psi_{i_{d_m},j_{d_m}} X^\upsilon)].$$

Additionally, for any $\theta \in \Theta^+$, we define the matrices V, W and IL_m as

$$\begin{cases} V := \mathrm{cov}_\theta(\chi_{m[0,0]}), \\ W_{[k,l]} := \dfrac{1}{p^2}\,\mathrm{tr}\big[C(\Psi_{i_k,j_k})\big(I_{p^2} - C(\theta_{m,\rho_1})\big)^2\big(I_{p^2} - C(\theta)\big)^{-2} C(\Psi_{i_l,j_l})\big] \\ \qquad \text{for any } k = 1, \dots, d_m, \\ IL_m := \mathrm{Diag}(\|\Psi_{i_k,j_k}\|_F^2, k = 1, \dots, d_m), \end{cases}$$

where for any vector u, $\mathrm{Diag}(u)$ is the diagonal matrix whose diagonal elements are the components of u. We also define the corresponding quantities $\chi^{\mathrm{iso}}_{m[0,0]}$, V^{iso}, W^{iso} and IL^{iso}_m in order to consider the isotropic estimator $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$.

PROPOSITION 4.4. *Let m be a model in $\mathcal{M}_1 \setminus \{\varnothing\}$, and let θ be an element of Θ^+_m that satisfies $(\mathbb{H}_1)$. Then $\widehat{\theta}_{m,\rho_1}$ converges to θ in probability, and*

$$\lim_{n \to +\infty} np^2 \mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)] = 2\sigma^4\,\mathrm{tr}[IL_m V^{-1}]. \tag{28}$$

Let θ in Θ^+ such that $(\mathbb{H}_2)$ is fulfilled. Then, $\widehat{\theta}_{m,\rho_1}$ converges to θ_{m,ρ_1} in probability and

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})] = 2\sigma^4\operatorname{tr}(WV^{-1}). \tag{29}$$

Both results still hold for the estimator $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ if θ belongs to $\Theta^{+,\mathrm{iso}}$ and if one replaces V, W, and IL_m by V^{iso}, W^{iso} and IL^{iso}_m.

In the first case, assumption $(\mathbb{H}_1)$ ensures that $\theta \in \Theta^+_{m,\rho_1}$ whereas assumption $(\mathbb{H}_2)$ ensures that $\theta_{m,\rho_1} \in \Theta^+_{m,\rho_1}$. The proof is based on the extension of Guyon's approach in the toroidal framework.

Expressions (28) and (29) are not easily interpretable in the present form. This is why we first derive (28) when θ is zero. Observe that it is equivalent to the independence of $(X_{[i,j]})_{(i,j)\in\Lambda}$.

EXAMPLE 4.5. *Assume that θ is zero. Then for any model $m \in \mathcal{M}_1$, the asymptotic risks of $\widehat{\theta}_{m,\rho_1}$ and $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ satisfy*

$$\lim_{n\to+\infty} np^2\mathbb{E}_{0_p}[l(\widehat{\theta}_{m,\rho_1},0_p)] = 2\sigma^2 d_m$$

and

$$\lim_{n\to+\infty} np^2\mathbb{E}_{0_p}[l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1},0_p)] = 2\sigma^2 d^{\mathrm{iso}}_m,$$

where we recall that d^{iso}_m is the dimension of the space Θ^{iso}_m.

PROOF. Since the components of X are independent, the matrix V equals $\sigma^2 IL_m$. We conclude by applying Proposition 4.4. □

Therefore, when the variables $X_{[i,j]}$ are independent, the asymptotic risk of $\widehat{\theta}_{m,\rho_1}$ equals, up to a factor 2, the variance term of the least squares estimator in the fixed design Gaussian regression framework. This quantity is of the same order as the penalty introduced in Section 3. When the matrix θ is nonzero, we can lower bound the limits (28) and (29).

COROLLARY 4.6. *Let m be a model in $\mathcal{M}_1$ and let $\theta \in \Theta^+_m$ that satisfies $(\mathbb{H}_1)$. Then, the variance term is asymptotically lower bounded as follows*:

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta)] \geq L\sigma^2\varphi_{\min}[I_{p^2} - C(\theta)]d_m = L\sigma^4\frac{d_m}{\varphi_{\max}(\Sigma)}, \tag{30}$$

where L is a universal constant. Let $\theta \in \Theta^+$ that satisfies $(\mathbb{H}_2)$. For any model $m \in \mathcal{M}_1$,

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})] \geq L\sigma^2(1-\|\theta\|_1)^3 d_m. \tag{31}$$

The proof is postponed to the technical Appendix [36]. Again, analogous lower bounds hold for $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ when θ belongs to $\Theta^{\mathrm{iso},+}$. This corollary states that asymptotically with respect to n the variance term of $\widehat{\theta}_{m,\rho_1}$ is larger than the order $d_m/(np^2)$. This expression is not really surprising since d_m stands for the dimension of the model m and np^2 corresponds to the number of data observed. Let us define $R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}) := \lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$ as the asymptotic variance term for $\widehat{\theta}_{m,\rho_1}$ rescaled by the number np^2 of observations.

The first part of corollary (30) states that from an asymptotic point of view the upper bound (25) is optimal. By Theorem 3.1, if we choose $\mathrm{pen}(m) = K\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$, then it holds that

$$\mathbb{E}[l(\widetilde{\theta}_{\rho_1},\theta)] \le L\bigl(K,\rho_1,\varphi_{\min}[I_{p^2}-C(\theta)]\bigr)\frac{R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta)}{np^2}$$

for any model $m\in\mathcal{M}\setminus\varnothing$ and any $\theta\in\Theta_m^+$ that satisfies $(\mathbb{H}_1)$. This property holds for any n and any p. Hence, $\widetilde{\theta}_{\rho_1}$ performs as well as the parametric estimator $\widehat{\theta}_{m,\rho_1}$ if the support of θ belongs to some unknown model m and if θ satisfies $(\mathbb{H}_1)$.

If we assume that $\|\theta\|_1<1$ [hypothesis $(\mathbb{H}_2)$], we are able to derive a stronger result.

PROPOSITION 4.7. *Considering $K\ge K_0$, $\rho_1\ge 2$, $\eta<1$ and a collection $\mathcal{M}\subset\mathcal{M}_1\setminus\varnothing$, we define the estimator $\widetilde{\theta}_{\rho_1}$ with the penalty* $\mathrm{pen}(m) = K\rho_1^2\frac{d_m}{np^2(1-\eta)}$. *Then the risk of $\widetilde{\theta}_{\rho_1}$ is upper bounded by*

$$\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1},\theta)] \le L(K,\rho_1,\eta)\inf_{m\in\mathcal{M}}\left\{l(\theta_{m,\rho_1},\theta)+\frac{R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})}{np^2}\right\} \tag{32}$$

for any $\theta\in\Theta^+\cap\mathcal{B}_1(0_p,\eta)$.

Observe that this property holds for any n and any p. If the matrix θ is strictly diagonally dominant, we therefore obtain an upper bound similar to an oracle inequality, except that the variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$ has been replaced by its asymptotic counterpart $R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})/(np^2)$. However, this inequality is not valid uniformly over any $\eta<1$: when η converges to one, the constant $L(K,\rho_1,\eta)$ tends to infinity. Indeed, if $\|\theta\|_1$ converges to one, the lower bound (31) on the variance term can behave like $(1-\|\theta\|_1)^3 d_m/(np^2)$ for some matrices θ whereas the penalty term $d_m/[np^2(1-\|\theta\|_1)]$ tends to infinity.

In the remaining part of the section, we illustrate that the constant $L(K,\eta,\rho_1)$ has to go to infinity when η goes to one. Let us consider the model m_1. It consists of GMRFs with 4-nearest neighbors.

EXAMPLE 4.8. *Let θ be a nonzero element of $\Theta^{\mathrm{iso}}_{m_1}$; then the asymptotic risk of $\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1}$ simplifies as*

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1},\theta)] = 2\frac{\sigma^4\theta_{[1,0]}}{\mathrm{cov}(X_{[1,0]},X_{[0,0]})}. \tag{33}$$

If we let size p of the network tend to infinity and $\theta_{[1,0]}$ go to $1/4$, the risk is equivalent to

$$\lim_{p\to+\infty}\lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1},\theta)] \underset{\theta_{[1,0]}\to 1/4}{\sim} \frac{16\sigma^2(1-4\theta_{[1,0]})}{\log(16)}.$$

The proof is postponed to the technical Appendix [36]. If follows from the second result that the lower bound (30) is sharp since in this particular case $\varphi_{\min}(I_{p^2}-C(\theta))=\sigma^2(1-4\theta_{[1,0]})$. When $\theta_{[1,0]}$ tends to $1/4$, then $\|\theta\|_1$ tends to one, and $\mathbb{E}_\theta[l(\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1},\theta)]$ behaves like $\sigma^2(1-\|\theta\|_1)d^{\mathrm{iso}}_{m_1}/(np^2)$ whereas the penalty $\mathrm{pen}(m_1)$ given in Theorem 3.1 has to be larger than $\sigma^2 d^{\mathrm{iso}}_{m_1}/[np^2(1-\|\theta\|_1)]$. Hence, the variance term and the penalty $\mathrm{pen}(\cdot)$ are not necessarily of the same order when $\|\theta\|_1$ tends to one. Theorem 3.1 cannot lead to an oracle inequality of the type (32) which is valid uniformly on $\eta<1$.

EXAMPLE 4.9. *Let α be a positive number smaller than $1/4$. For any integer p which is divisible by 4, we define the $p\times p$ matrix $\theta^{(p)}$ by*

$$\begin{cases}\theta^{(p)}_{[p/4,p/4]}=\theta^{(p)}_{[-p/4,p/4]}=\theta^{(p)}_{[p/4,-p/4]}=\theta^{(p)}_{[-p/4,-p/4]}:=\alpha,\\ \theta^{(p)}_{[i,j]}:=0, \qquad \textit{else}.\end{cases}$$

Then the variance term is asymptotically lower bounded as follows:

$$\lim_{p\to+\infty}\lim_{n\to+\infty} np^2\mathbb{E}_{\theta^{(p)}}\big[l\big(\widehat{\theta^{(p)}}^{\mathrm{iso}}_{m_1,\rho_1},[\theta^{(p)}]^{\mathrm{iso}}_{m_1,\rho_1}\big)\big]\geq \frac{L\sigma^2}{1-4\alpha}.$$

The proof is postponed to the technical Appendix [36]. This variance term is of order $\sigma^2 d^{\mathrm{iso}}_m/[np^2(1-\|\theta\|_1)]=\varphi_{\max}(\Sigma)d^{\mathrm{iso}}_m/(np^2)$ when $\|\theta\|_1$ goes to one. The penalty $\mathrm{pen}(m)$ introduced in Proposition 4.7 is therefore a sharp upper bound of the variance terms.

On one hand, we take a penalty $\mathrm{pen}(m)$ larger than $\sigma^2 d_m/(np^2(1-\|\theta\|_1))$. On the other hand, the variance of $\widehat{\theta}_{m,\rho_1}$ is of the order $\sigma^2(1-\|\theta\|_1)d_m/(np^2)$ in some cases. Bound (32) cannot, therefore, hold uniformly over any $\eta<1$. We think that it is intrinsic to the penalization strategy.

5. Comments on the assumptions. In this section, we discuss the dependency of the estimators $\widehat{\theta}_{m,\rho_1}$ on ρ_1 as well as assumptions $(\mathbb{H}_1)$ and $(\mathbb{H}_2)$.

Dependency of $\widehat{\theta}_{m,\rho_1}$ on ρ_1. We recall that the estimator $\widehat{\theta}_{m,\rho_1}$ is defined in (18) as the minimizer of the CLS empirical contrast $\gamma_{n,p}(\cdot)$ over Θ^+_{m,ρ_1}. It may seem restrictive to perform the minimization over the set Θ^+_{m,ρ_1} instead of Θ^+_m. Nevertheless, we advocate that it is not the case, at least for small models. Let us indeed define

$$\rho(m):=\sup_{\theta\in\Theta^+_m}\varphi_{\max}[I_{p^2}-C(\theta)] \quad\text{and}\quad \rho^{\mathrm{iso}}(m):=\sup_{\theta\in\Theta^{+,\mathrm{iso}}_m}\varphi_{\max}[I_{p^2}-C(\theta)].$$

TABLE 2
Approximate computation of $\rho(m)$ and $\rho^{\mathrm{iso}}(m)$ for the four smallest models with $p = 50$

d_m	2	4	6	10
$\rho(m)$	2.0	4.0	5.0	6.8
d_m^{iso}	1	2	3	4
$\rho^{\mathrm{iso}}(m)$	2.0	4.0	5.0	6.8

The quantities $\rho(m)$ and $\rho^{\mathrm{iso}}(m)$ are finite since Θ_m^+ is bounded. If one takes ρ_1 larger than $\rho(m)$ [resp., $\rho^{\mathrm{iso}}(m)$], then the set Θ_{m,ρ_1}^+ (resp., $\Theta_{m,\rho_1}^{+,\mathrm{iso}}$) is exactly Θ_m^+ (resp., $\Theta_m^{+,\mathrm{iso}}$). We illustrate in Table 2 that $\rho(m)$ and $\rho^{\mathrm{iso}}(m)$ are small when model m is small. Consequently, choosing a moderate value for ρ_1 is not really restrictive for small models. However, when the size of model m increases, the sets Θ_{m,ρ_1}^+ and Θ_m^+ become different for moderate values of ρ_1. In Section 7, we discuss the choice of ρ_1.

Assumption $(\mathbb{H}_1)$ defined in (22) states that the largest eigenvalue of $(I_{p^2} - C(\theta))$ is smaller than ρ_1. We have illustrated in Table 2 that if the support of θ belongs to a small model m, then the maximal absolute value of $(I_{p^2} - C(\theta))$ is small. Hence, assumption $(\mathbb{H}_1)$ is ensured for "moderate" values of ρ_1 as soon as the support of θ belongs to some small model. If θ is not sparse but approximately sparse it is likely that the largest eigenvalue of θ remain moderate. In practice, we do not know in advance if a given choice of ρ_1 ensures $(\mathbb{H}_1)$. In Section 7, we discuss an extension of our procedure which does not require assumption $(\mathbb{H}_1)$.

Assumption $(\mathbb{H}_2)$ defined in (27) states that $\theta \in \mathcal{B}_1(0_p, 1)$ or equivalently that the matrix $(I_{p^2} - C(\theta))$ is diagonally dominant. Rue and Held prove in [29], Section 2.7, that $\Theta_{m_1}^+$ is included in $\mathcal{B}_1(0_p, 1)$. They also point out that a small part of $\Theta_{m_2}^+$ does not belong to $\mathcal{B}_1(0_p, 1)$. In fact, assumption $(\mathbb{H}_2)$ becomes more and more restrictive if the support of θ becomes larger. Nevertheless, assumption $(\mathbb{H}_2)$ is also quite common in the literature (as, for instance, in [17]).

If one looks closely at our proofs involving assumption $(\mathbb{H}_2)$, one realizes that this assumption is only made to ensure the following facts:

1. The *projection* θ_{m,ρ_1} belongs to the open set Θ_{m,ρ_1}^+ for any model $m \in \mathcal{M}$ (Corollary 4.3).
2. The smallest eigenvalue of $(I_{p^2} - C(\theta_{m,\rho_1}))$ is lower bounded by some positive number ρ_2, uniformly over all models $m \in \mathcal{M}$.

From empirical observations, these two last facts seem far more restrictive than $(\mathbb{H}_2)$. We used assumption $(\mathbb{H}_2)$ in the statement of our results, because we did not find any weaker but still simple condition that ensures facts 1 and 2.

6. Minimax rates. In Theorem 3.1 and Proposition 4.7 we have shown that under mild assumptions on θ the estimator $\widetilde{\theta}_{\rho_1}$ behaves almost as well as the best estimator among the family $\{\widehat{\theta}_{m,\rho_1}, m \in \mathcal{M}\}$. We now compare the risk of $\widetilde{\theta}_{\rho_1}$ with the risk of any other possible estimator $\widehat{\theta}$. This includes comparison with maximum likelihood methods. There is no hope to make a pointwise comparison with an arbitrary estimator. Therefore, we classically consider the maximal risk over some suitable subsets $\mathcal{T}$ of Θ^+. The *minimax risk* over the set $\mathcal{T}$ is given by $\inf_{\widehat{\theta}} \sup_{\theta\in\mathcal{T}} \mathbb{E}_\theta[l(\widehat{\theta},\theta)]$ where the infimum is taken over all possible estimators $\widehat{\theta}$ of θ. Then the estimator $\widetilde{\theta}_{\rho_1}$ is said to be *approximately minimax* with respect to the set $\mathcal{T}$ if the ratio,

$$\frac{\sup_{\theta\in\mathcal{T}} \mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1},\theta)]}{\inf_{\widehat{\theta}} \sup_{\theta\in\mathcal{T}} \mathbb{E}_\theta[l(\widehat{\theta},\theta)]},$$

is smaller than a constant that does not depend on σ^2, n or p. An estimator is said to be *adaptive* to a collection $(\mathcal{T}_i)_{i\in\mathcal{I}}$ if it is simultaneously minimax over each $\mathcal{T}_i$. The problem of designing adaptive estimation procedures is in general difficult. It has been extensively studied in the fixed design Gaussian regression framework. See for instance [6] for a detailed discussion. In the sequel, we adapt some of their ideas to the GMRF framework.

We prove in Section 6.1 that the estimator $\widetilde{\theta}_{\rho_1}$ is adaptive to the unknown sparsity of the matrix θ. Moreover, it is also adaptive if we consider the Frobenius distance between partial correlation matrices. In Section 6.2, we show that $\widetilde{\theta}_{\rho_1}$ is also adaptive to the rates of decay of the bias.

We need to restrain ourselves to set of matrices θ such that the largest eigenvalue of the covariance matrix Σ is uniformly bounded. This is why we define

$$\forall \rho_2 > 1 \qquad \mathcal{U}(\rho_2) := \left\{\theta \in \Theta, \varphi_{\min}\big(I_{p^2} - C(\theta)\big) \geq \frac{1}{\rho_2}\right\}. \tag{34}$$

Observe that $\theta \in \mathcal{U}(\rho_2)$ is exactly equivalent to $\varphi_{\max}(\Sigma) \leq \sigma^2\rho_2$ since $\Sigma = \sigma^2(I_{p^2} - C(\theta))$.

6.1. *Adapting to unknown sparsity.* In this subsection, we prove that under mild assumptions the penalized estimator $\widetilde{\theta}_{\rho_1}$ is adaptive to the unknown sparsity of θ. We first lower bound the minimax rate of convergence on given hypercubes.

DEFINITION 6.1. Let m be a model in the collection $\mathcal{M}_1 \setminus \varnothing$. We consider $(\Psi_{i_1,j_1}, \ldots, \Psi_{i_{d_m},j_{d_m}})$ a basis of the space Θ_m defined by (14). For any $\theta' \in \Theta_m^+$, the hypercube $\mathcal{C}_m(\theta', r)$ is defined as

$$\mathcal{C}_m(\theta', r) := \left\{\theta' + \sum_{k=1}^{d_m} \Psi_{i_k,j_k}\phi_k, \phi \in \{0,1\}^{d_m}\right\},$$

if the positive number r is small enough so that $\mathcal{C}_m(\theta', r) \subset \Theta^+$. For any $\theta' \in \Theta_m^{+,\mathrm{iso}}$, we analogously define the hypercubes $\mathcal{C}_m^{\mathrm{iso}}(\theta', r)$ using a basis $(\Psi^{\mathrm{iso}}_{i_1,j_1}, \ldots, \Psi^{\mathrm{iso}}_{i_{d_m},j_{d_m}})$.

PROPOSITION 6.2. *Let m be a model in $\mathcal{M}_1 \setminus \varnothing$ whose dimension d_m is smaller than $p\sqrt{n}$. Then, for any estimator $\widehat{\theta}$,*

$$\sup_{\theta \in \Theta_m^+} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] \geq \sup_{\theta \in \Theta_{m,2}^+} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] \geq L\sigma^2 \frac{d_m}{np^2}. \tag{35}$$

Let θ' be an element of Θ_m^+ that satisfies $(\mathbb{H}_2)$. For any estimator $\widehat{\theta}$ of θ,

$$\sup_{\theta \in \mathrm{Co}[\mathcal{C}_m(\theta', (1-\|\theta'\|_1)/\sqrt{np^2})]} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] \geq L\sigma^2 \varphi_{\min}^2[I_{p^2} - C(\theta')] \frac{d_m}{np^2}, \tag{36}$$

where $\mathrm{Co}[\mathcal{C}_m(\theta', r)]$ denotes the convex hull of $\mathcal{C}_m(\theta', r)$.

An analogous result holds for isotropic hypercubes. The first bound (35) means that for any estimator $\widehat{\theta}$, the supremum of the risks $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ over Θ_m^+ is larger than $\sigma^2 d_m/(np^2)$ (up to some numerical constant). This rate $\sigma^2 d_m/(np^2)$ is achieved by the CLS estimator by Theorem 3.1.

The second lower bound (36) is of independent interest. It implies that in a small neighborhood of θ' the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ is larger than $\sigma^2 \varphi_{\min}^2[I_{p^2} - C(\theta')] d_m/(np^2)$. This confirms the lower bound (30) of Corollary 4.6 in a nonasymptotic way. Indeed, these two expressions match up to a factor $\varphi_{\min}[I_{p^2} - C(\theta')]$. This difference comes from the fact that the lower bound (36) holds for any estimator $\widehat{\theta}$. Bound (36) is sharp in the sense that the maximum likelihood estimator $\widehat{\theta}^{\mathrm{iso,mle}}_{m_1}$ of isotropic GMRF in m_1 exhibits an asymptotic risk of order $\sigma^2 \varphi_{\min}^2[I_{p^2} - C(\theta)]/(np^2)$ for the parameter θ studied in Example 4.8. It is shown using the methodology introduced in the proof of Example 4.8. We now state that $\widetilde{\theta}_\rho$ is adaptive to the sparsity of m.

COROLLARY 6.3. *Considering $K \geq K_0$, $\rho_1 \geq 2$, $\rho_2 > 2$ and a collection $\mathcal{M} \subset \mathcal{M}_1$, we define the estimator $\widetilde{\theta}_{\rho_1}$ with the penalty* $\mathrm{pen}(m) = K\sigma^2 \rho_1^2 \rho_2 \frac{d_m}{np^2}$. *For any nonempty model m,*

$$\sup_{\theta \in \Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] \leq L(K, \rho_1, \rho_2) \inf_{\widehat{\theta}} \sup_{\theta \in \Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)} \mathbb{E}[l(\widehat{\theta}, \theta)], \tag{37}$$

where $\mathcal{U}(\rho_2)$ is defined in (34).

A similar result holds for $\widetilde{\theta}^{\mathrm{iso}}_{\rho_1}$ and $\Theta^{+,\mathrm{iso}}_{m,\rho_1}$. Corollary 6.3 is nonasymptotic and applies for any n and any p. If θ belongs to some model m, then the optimal risk from a minimax point of view is of order $\frac{d_m}{np^2}$. In practice, we do not know the

true model m. Nevertheless, the procedure simultaneously achieves the minimax rates for all supports m possible. This means that $\widetilde{\theta}_{\rho_1}$ reaches this minimax rate $\frac{d_m}{np^2}$ *without* knowing in advance the true model m.

The procedure is not adaptive to the smallest or the largest eigenvalue of $(I_{p^2} - C(\theta))$ which correspond to ρ_1 and ρ_2. Indeed, the constant $L(K, \rho_1, \rho_2)$ depends on ρ_1 and ρ_2. We are not aware of any other covariance estimation procedure which is really adaptive to the smallest or the largest eigenvalue of the matrix.

Finally, $\widetilde{\theta}_{\rho_1}$ exhibits the same adaptive properties with respect to the Frobenius norm.

COROLLARY 6.4. *Under the same assumptions as Corollary 6.3,*

$$\sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \mathbb{E}_\theta[\|C(\widetilde{\theta}_{\rho_1}) - C(\theta)\|_F^2] \le L(K,\rho_1,\rho_2)\inf_{\widehat{\theta}} \sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \mathbb{E}[\|C(\widehat{\theta}) - C(\theta)\|_F^2].$$

PROOF. As in the proof of Corollary 3.2, we observe that

$$\|C(\theta_1) - C(\theta_2)\|_F \ge \frac{p^2\rho_1}{\sigma^2} l(\theta_1,\theta_2),$$

if θ satisfies assumption $(\mathbb{H}_1)$. We conclude by applying Proposition 6.2 and Corollary 3.2. □

6.2. *Adapting to the decay of the bias.* In this section, we prove that the estimator $\widetilde{\theta}_{\rho_1}$ is adaptive to a range of sets that we call *pseudo-ellipsoids*.

DEFINITION 6.5 (Pseudo-ellipsoids). Let $(a_j)_{1\le j\le \mathrm{Card}(\mathcal{M}_1)}$ be a nonincreasing sequence of positive numbers. Then, $\theta \in \Theta^+$ belongs to the *pseudo-ellipsoid* $\mathcal{E}(a)$ if and only if

$$\sum_{i=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{\mathrm{var}_\theta(X_{[0,0]}|X_{\mathcal{N}(m_{i-1})}) - \mathrm{var}_\theta(X_{[0,0]}|X_{\mathcal{N}(m_i)})}{a_i^2} \le 1. \tag{38}$$

Condition (38) measures how fast $\mathrm{Var}_\theta(X_{[0,0]}|X_{\mathcal{N}(m_i)})$ tends to $\mathrm{Var}_\theta(X_{[0,0]}|X_{\Lambda\setminus\{(0,0)\}})$. Suppose that assumption $(\mathbb{H}_2)$ defined in (27) is fulfilled. By Corollary 4.2, $\mathrm{Var}_\theta(X_{[0,0]}|X_{\mathcal{N}(m_i)})$ is the sum of $l(\theta_{m_i},\theta)$ and σ^2, and condition (38) is equivalent to

$$\sum_{i=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{l(\theta_{m_{i-1}},\theta) - l(\theta_{m_i},\theta)}{a_i^2} \le 1. \tag{39}$$

Hence, the sequence (a_i) gives some condition on the *rate of decay* of the bias when the dimension of the model increases. These sets $\mathcal{E}(a)$ are not true ellipsoids. Nevertheless, one may consider them as counterparts of the classical ellipsoids studied in the fixed design Gaussian regression framework (see, for instance, [25], Section 4.3).

To prove adaptivity, we shall need the equivalence between conditions (38) and (39). This equivalence holds if $\mathrm{Var}_\theta(X_{[0,0]}|X_{\mathcal{N}(m_i)})$ decomposes as $l(\theta_{m_i},\theta)+\sigma^2$ for any model $m\in\mathcal{M}_1$. As mentioned earlier, assumption $(\mathbb{H}_2)$ is sufficient (but not necessary) for this property to hold. This is why we restrict ourselves to study sets of the type $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)$. We shall also perform the following assumption on the ellipsoids $\mathcal{E}(a)$:

$$(\mathbb{H}_a):\qquad a_i^2\le\frac{\sigma^2}{d_{m_i}}\qquad\text{for any } 1\le i\le|\mathcal{M}_1|.$$

It essentially means that the sequence (a_i) converges fast enough toward 0. For instance, all the sequences $a_i=\sigma(d_{m_i})^{-s}$ with $s\ge 1/2$ satisfy $(\mathbb{H}_a)$.

PROPOSITION 6.6. *Under assumption* $(\mathbb{H}_a)$, *the minimax rate of estimation on* $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)$ *is lower bounded by*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)}\mathbb{E}_\theta[l(\widehat{\theta},\theta)]\ge L\sup_{1\le i\le\mathrm{Card}(\mathcal{M}_1)}\biggl(a_i^2\wedge\sigma^2\frac{d_{m_i}}{np^2}\biggr).\tag{40}$$

This lower bound is analogous to the minimax rate of estimation for ellipsoids in the Gaussian sequence model. Gathering Theorem 3.1 and Proposition 6.6 enables to derive adaptive properties for $\widetilde{\theta}_{\rho_1}$.

PROPOSITION 6.7. *Considering* $K\ge K_0$, $\rho_1\ge 2$, $\rho_2>2$ *and the collection* $\mathcal{M}_1$, *we define the estimator* $\widetilde{\theta}_{\rho_1}$ *with the penalty* $\mathrm{pen}(m)=K\sigma^2\rho_1^2\rho_2\frac{d_m}{np^2}$. *For any ellipsoid* $\mathcal{E}(a)$ *that satisfies* $(\mathbb{H}_a)$ *and such that* $a_1^2\ge 1/(np^2)$, *the estimator* $\widetilde{\theta}_{\rho_1}$ *is minimax over the set* $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)$,

$$\begin{aligned}&\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1},\theta)]\\&\quad\le L(K,\rho_1,\rho_2)\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta[l(\widehat{\theta},\theta)].\end{aligned}\tag{41}$$

Let us first illustrate this result. We have mentioned earlier that assumption $(\mathbb{H}_a)$ is satisfied for all sequences $a_i=\sigma(d_{m_i})^{-s}$ with $s\ge 1/2$. We note that $\mathcal{E}'(s)$ such a pseudo-ellipsoid. By Propositions 6.6 and 6.7, the minimax rate over *one* pseudo ellipsoid $\mathcal{E}'(s)$ is $\sigma^2(np^2)^{-2s/(1+2s)}$. The larger s is, the faster the minimax rates is. The estimator $\widetilde{\theta}_{\rho_1}$ achieves simultaneously the rate $\sigma^2(np^2)^{-2s/(1+2s)}$ for all

$s \geq 1/2$. Consequently, $\widetilde{\theta}_{\rho_1}$ is adaptive to the rate s of decay of the bias: it achieves the optimal rates without knowing s in advance.

Let us further comment on Proposition 6.7. By (41), the estimator $\widetilde{\theta}_{\rho_1}$ is adaptive over $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(\rho_2)$ for all sequences (a) such that $(\mathbb{H}_a)$ is satisfied and such that $a_1^2 \geq 1/(np^2)$. Again, the result applies for any n and any p. The condition $a_1^2 \geq 1/(np^2)$ is classical. It ensures that the pseudo-ellipsoid $\mathcal{E}(a)$ is not degenerate, that is, that the minimax rates of estimation is not smaller than $\sigma^2/(np^2)$. We explained earlier that we restrict ourselves to parameters θ in $\mathcal{B}_1(0_p, 1)$ only because this enforces the equivalence between (38) and (39). In contrast, the hypothesis $\varphi_{\max}(\Sigma) \leq \sigma^2 \rho_2$ is really necessary because we fail to be adaptive to ρ_2.

COROLLARY 6.8. *Under assumption* $(\mathbb{H}_a)$, *the minimax rate of estimation over* $\mathcal{E}(a) \cap \mathcal{U}(2) \cap \mathcal{B}_1(0_p, 1)$ *is lower bounded by*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(2)} \mathbb{E}_\theta[\|C(\widehat{\theta}) - C(\theta)\|_F^2] \geq L \sup_{1 \leq i \leq \mathrm{Card}(\mathcal{M}_1)} \left(a_i^2 p^2 \wedge \frac{d_{m_i}}{n} \right).$$

Under the same assumptions as Proposition 6.7,

$$\sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta[\|C(\widehat{\theta}) - C(\theta)\|_F^2] \leq L(K, \rho_1, \rho_2) \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta[\|C(\widehat{\theta}) - C(\theta)\|_F^2].$$

PROOF. As in the proof of Corollary 3.2, we observe that

$$\|C(\theta_1) - C(\theta_2)\|_F \geq p^2 [\varphi_{\max}(\Sigma)]^{-1} l(\theta_1, \theta_2) \geq \frac{p^2}{\rho_2 \sigma^2} l(\theta_1, \theta_2),$$

$$\|C(\theta_1) - C(\theta_2)\|_F \leq p^2 [\varphi_{\min}(\Sigma)]^{-1} l(\theta_1, \theta_2) \leq p^2 \frac{\varphi_{\max}[I_{p^2} - C(\theta)]}{\sigma^2} l(\theta_1, \theta_2) \leq \frac{\rho_2 p^2}{\sigma^2} l(\theta_1, \theta_2),$$

if $\theta \in \mathcal{B}_1(0_p, 1) \cap \mathcal{B}_{\mathrm{op}}(\rho_2)$. We conclude by applying Propositions 6.6 and 6.7. □

Again, $\widetilde{\theta}_{\rho_1}$ satisfies the same minimax properties with respect to the Frobenius norm. All these properties easily extend to isotropic fields if one defines the corresponding sets $\mathcal{E}^{\mathrm{iso}}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(\rho_2)$ of isotropic GMRFs.

7. Discussion.

7.1. *Comparison with maximum likelihood estimation.* Let us first compare the computational cost, the CLS estimation method and the maximum likelihood

estimator (MLE). For toroidal lattices, fast algorithms based on two-dimensional fast-Fourier transformation (see, for instance, [31]) allow to compute the MLE as fast as the CLS estimator. More details on the computation of the CLS estimators for toroidal lattices are given in [37], Section 2.3. When the lattice is not a torus, the MLE becomes intractable because it involves the optimization of a determinant of size p^2. In contrast, the CLS criterion $\gamma_{n,p}(\cdot)$ defined in (16) is a quadratic function of θ. Consequently, CLS estimators are still computationally amenable. We extend our model selection to nontoroidal lattices in [37].

Let us compare the risk of CLS estimators and MLE. Given a small-dimensional model m, the risk of the *parametric* CLS estimator and the *parametric* MLE have been compared from an asymptotic point of view ([17], Section 4.3). It is generally accepted (see, for instance, Cressie [10], Section 7.3.1) and that *parametric* CLS estimators are almost as efficient as parametric MLE for the major part of the parameter spaces Θ_m^+. We have nonasymptotically assessed this statement in Proposition 6.2 by minimax arguments. Nevertheless, for some parameters θ that are close to the border of Θ_m^+, Kashyap and Chellappa [22] have pointed out that CLS estimators are less efficient than MLE. If we have proved nonasymptotic bounds for CLS-based model selection method, we are not aware of any such result for model selection procedures based on MLE.

7.2. *Concluding remarks*. We have developed a model selection procedure for choosing the neighborhood of a GMRF. In Theorem 3.1, we have proven a nonasymptotic upper bound for the risk of the estimator $\widetilde{\theta}_{\rho_1}$ with respect to the prediction error $l(\cdot,\cdot)$. Under assumption $(\mathbb{H}_1)$, this bound is shown to be optimal from an asymptotic point of view if the support of θ belongs to one of the models in the collection. If assumption $(\mathbb{H}_2)$ is fulfilled, we are able to obtain an oracle-type inequality for $\widetilde{\theta}_{\rho_1}$. Moreover, $\widetilde{\theta}_{\rho_1}$ is minimax adaptive to the sparsity of θ under $(\mathbb{H}_1)$. Finally, it simultaneously achieves the minimax rates of estimation over a large class of sets $\mathcal{E}(a)$ if $(\mathbb{H}_2)$ holds. Some of these properties still hold if we use the Frobenius loss function. The case of isotropic Gaussian fields is handled similarly.

However, in the oracle inequality (32) and in the minimax bounds (37) and (41), we either perform an assumption on the l_1 norm of θ or on the smallest eigenvalue of $(I_{p^2} - C(\theta))$. When $\|\theta\|_1$ tends to one or $\varphi_{\min}[I_{p^2} - C(\theta)]$ tends to 0, there is a distortion between the upper bound $\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)]$ provided by Theorem 3.1 and the lower bounds given by Corollary 4.6 or Proposition 6.2. This limitation seems intrinsic to our penalization method which is linear with respect to the dimension, whereas the asymptotic variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ depends in a complex way on the dimension of the model m and on the target θ. In our opinion, achieving adaptivity with respect to the smallest eigenvalue of $(I_{p^2} - C(\theta))$ (or, equivalently, the largest value of Σ) would require a different penalization technique. Nevertheless, we are not aware of any procedure in a covariance estimation setting that is adaptive to the largest eigenvalues of Σ.

So far, we have provided an estimation procedure for $(I_{p^2} - C(\theta)) = \sigma^2 \Sigma^{-1}$. If we aim at estimating the precision matrix Σ^{-1}, we also have to take into account the quantity σ^2. It is natural to estimate it by $\widetilde{\sigma}^2 := \gamma_{n,p^2}(\widetilde{\theta}_{\rho_1})$ as done for instance by Guyon in [17], Section 4.3, in the parametric setting. Then, we obtain the estimate $\widetilde{\Sigma}^{-1} := \widetilde{\sigma}^2(I_{p^2} - C(\widetilde{\theta}_{\rho_1})$. It is of interest to study the adaptive properties of this estimator with respect to loss functions such as the Frobenius or operator norm as is done in [28] in the nonstationary setting. Nevertheless, let us mention that the matrix $\widetilde{\Sigma}^{-1}$ is not necessarily invertible since the estimator $\widetilde{\theta}_{\rho_1}$ belongs to the closure of Θ^+.

The choice of the quantity ρ_1 is problematic. On one hand, ρ_1 should be large enough so that assumption $(\mathbb{H}_1)$ is fulfilled. On the other hand, a large value of ρ_1 yields worse bounds in Theorem 3.1. Moreover, the largest eigenvalue of $(I_{p^2} - C(\theta))$ is unknown in practice, which makes more difficult the choice of ρ_1. We see two possible answers to this issue:

- First, moderate values of ρ_1 are sufficient to enforce $(\mathbb{H}_1)$ if the target θ is sparse as illustrated in Table 2.
- Second, we believe that the bounds for the risk are pessimistic with respect to ρ_1. A future direction of research is to derive risk bounds for $\widetilde{\theta}_{\rho_1}$ with $\rho_1 = +\infty$. In [37], we illustrate that such a procedure gives rather good results in practice.

In Theorem 3.1, we only provide a lower bound of the penalty so that the procedure performs well. However, this bound depends on the largest eigenvalue of Σ which is seldom known in practice and we did not give any advice for choosing a "reasonable" constant K in practice. This is why we introduce in [37] a data-driven method based on the *slope heuristics* of Birgé and Massart [7] for calibrating the penalty. We also provide numerical evidence of its performances on simulated data. For instance, the procedure outperforms variogram-based methods for estimating Matérn correlations.

We have mentioned in the Introduction that the toroidal assumption for the lattice is somewhat artificial in several applications. Nevertheless, we needed to neglect the edge effects in order to derive nonasymptotic properties for $\widetilde{\theta}_{\rho_1}$ as in Theorem 3.1. In practice, it is often more realistic to suppose that we observe a small window of a Gaussian field defined on the whole plane $\mathbb{Z}^2$. The previous nonasymptotic properties do not extend to this new setting. Nevertheless, Lakshman and Derin have shown in [23] that there is no phase transition within the valid parameter space for GMRFs defined on the plane $\mathbb{Z}^2$. In short, this implies that the distribution of a field observed in a fixed window of a GMRF does not asymptotically depend on the bound condition. Therefore, it is reasonable to think that our estimation procedure performs well if it was adapted to this new setting. In [37], we describe such an extension and we provide numerical evidence of its performances.

7.3. *Possible extensions.* In many statistical applications stationary Gaussian fields (or Gaussian Markov random fields) are not directly observed. For instance,

Aykroyd [1] or Dass and Nair [13] use compound Gaussian Markov random fields to account for nonstationarity and steep variations. The wavelet transform has emerged as a powerful tool in image analysis. the wavelet coefficients of an image are sometimes modeled using hidden Markov models [12, 27]. More generally, the success of the GMRF is mainly due to the use of hierarchical models involving latent GMRFs [30]. The study and the implementation of our penalization strategy for selecting the complexity of the latent Markov models is an interesting direction of research.

8. Proofs.

8.1. *A concentration inequality.* In this section, we prove a new concentration inequality for suprema of Gaussian chaos of order 2. It will be useful for proving Theorem 3.1.

PROPOSITION 8.1. *Let F be a compact set of symmetric matrices of size r, $(Y^1, \dots, Y^n)$ be a n-sample of a standard Gaussian vector of size r and Z be the random variable defined by*

$$Z := \sup_{R \in F} \operatorname{tr}[R(\overline{YY^*} - I_r)].$$

Then

$$\mathbb{P}\big(Z \geq \mathbb{E}(Z) + t\big) \leq \exp\left[-\left(\frac{t^2}{L_1 \mathbb{E}(W)} \wedge \frac{t}{L_2 B}\right)\right], \tag{42}$$

where the quantities B and W are such that

$$B := \frac{2}{n} \sup_{R \in F} \varphi_{\max}(R),$$

$$W := \frac{4}{n} \sup_{R \in F} \operatorname{tr}(R\overline{YY^*}R').$$

The main argument of this proof is to transfer a deviation inequality for suprema of Rademacher chaos of order 2 to suprema of Gaussian chaos. Talagrand [35] has first given in Theorem 1.2 a concentration inequality for such suprema of Rademacher chaos. Boucheron et al. [8] have recovered the upper bound applying a new methodology based on the entropy method. We adapt their proof to consider nonnecessarily homogeneous chaos of order 2. More details are found in the technical Appendix [36].

8.2. *Proof of Theorem 3.1.*

PROOF. We only consider the case of anisotropic estimators. The proofs and lemma are analogous for isotropic estimators. We first fix a model $m \in \mathcal{M}$. By

definition, the model $\widehat{m}$ satisfies

$$\gamma_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(\widehat{m}) \le \gamma_{n,p}(\theta_{m,\rho_1}) + \mathrm{pen}(m).$$

For any $\theta' \in \Theta^+$, $\overline{\gamma}_{n,p}(\theta')$ stands for the difference between $\gamma_{n,p}(\theta')$ and its expectation $\gamma(\theta')$. Then, the previous inequality turns into

$$\gamma(\widetilde{\theta}_{\rho_1}) \le \gamma(\theta_{m,\rho_1}) + \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m}).$$

Subtracting the quantity $\gamma(\theta)$ to both sides of this inequality yields

$$l(\widetilde{\theta}_{\rho_1}, \theta) \le l(\theta_{m,\rho_1}, \theta) + \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m}). \tag{43}$$

The proof is based on the control of the random variable $\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1})$.

LEMMA 8.2. *For any positive number α, ξ, and $\delta > 1$ the event Ω_ξ defined by*

$$\Omega_\xi = \left\{ \begin{array}{l} \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \\ \qquad \le \dfrac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \dfrac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) \\ \qquad\quad + \dfrac{K_0 \delta^2 \rho_1^2 \varphi_{\max}(\Sigma)}{np^2} \left[(1+\alpha/2)(d_m + d_{\widehat{m}}) + \dfrac{\xi^2}{\delta - 1} \right] \end{array} \right\},$$

satisfies

$$\mathbb{P}(\Omega_\xi^c) \le \exp\left\{ -L_1 \xi \left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n} \right] \right\} \times \sum_{m' \in \mathcal{M}} \exp\left\{ -L_2 \sqrt{d_{m'}} \left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2} \right) \right\}.$$

A similar lemma holds in the isotropic case. In particular, we choose $\alpha = (K - K_0)/K_0$ and $\delta = \sqrt{(1+\alpha)/(1+\alpha/2)}$. Lemma 8.2 implies that on the event Ω_ξ,

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \le \frac{1}{\sqrt{\delta(\alpha)}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta(\alpha)}}{\sqrt{\delta(\alpha)} - 1} l(\theta_{m,\rho_1}, \theta) + \mathrm{pen}(m) + \mathrm{pen}(\widehat{m}) + \frac{K_0 \xi^2 \delta(\alpha)^2 \rho_1^2 \varphi_{\max}(\Sigma)}{np^2(\delta(\alpha) - 1)}.$$

Thus, gathering this bound with inequality (43) yields

$$\frac{\delta(\alpha)^{1/2} - 1}{\delta(\alpha)^{1/2}} l(\widetilde{\theta}_{\rho_1}, \theta) \le \left[1 + \delta(\alpha)^{-1/2} \left(\delta(\alpha)^{1/2} - 1 \right)^{-1} \right] l(\theta_{m,\rho_1}, \theta) + 2\,\mathrm{pen}(m) + \frac{K_0 \xi^2 \rho_1^2 \varphi_{\max}(\Sigma) \delta(\alpha)^2}{np^2(\delta(\alpha) - 1)}$$

with probability larger than $1 - \mathbb{P}(\Omega_\xi)$. Integrating this inequality with respect to $\xi > 0$ leads to

$$\begin{aligned}
&\frac{\delta(\alpha)^{1/2} - 1}{\delta(\alpha)^{1/2}} \mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] \\
&\quad \le \bigl[1 + \delta(\alpha)^{-1/2}\bigl(\delta(\alpha)^{1/2} - 1\bigr)^{-1}\bigr] l(\theta_{m,\rho_1}, \theta) \\
&\qquad + 2\,\mathrm{pen}(m) + \frac{\delta(\alpha)^2 L(\alpha)}{(\delta(\alpha) - 1)[\alpha^2/(1 + \alpha/2) \wedge n]} \frac{\rho_1^2 \varphi_{\max}(\Sigma)}{np^2}.
\end{aligned} \tag{44}$$

We upper bound $[(\alpha^2/(1 + \alpha/2)) \wedge n]^{-1}$ by $[(\alpha^2/(1 + \alpha/2)) \wedge 1]^{-1}$. Since $\alpha = \frac{K - K_0}{K_0}$, it follows that

$$\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] \le L_1(K)[l(\theta_{m,\rho_1}, \theta) + \mathrm{pen}(m)] + L_2(K) \frac{\rho_1^2 \varphi_{\max}(\Sigma)}{np^2}.$$

Taking the infimum over the models $m \in \mathcal{M}$ allows us to conclude. □

PROOF OF LEMMA 8.2. Throughout this proof, it is more convenient to express the quantities $\overline{\gamma}_{n,p}(\cdot)$ and $l(\cdot)$ in terms of covariance and precision matrices. Thanks to (19), we also provide a matricial expression for $\gamma(\cdot)$:

$$\gamma(\theta') = \frac{1}{p^2} \mathrm{tr}\bigl[\bigl(I - C(\theta')\bigr)\Sigma\bigl(I - C(\theta')\bigr)\bigr]. \tag{45}$$

Gathering identities (45) and (17), we get

$$\begin{aligned}
&\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \\
&\quad = \frac{1}{p^2} \mathrm{tr}\bigl[\bigl([I_{p^2} - C(\theta_{m,\rho_1})]^2 - [I_{p^2} - C(\widetilde{\theta}_{\rho_1})]^2\bigr)\bigl(\overline{\mathbf{X}^{\mathbf{v}}\mathbf{X}^{\mathbf{v}*}} - \Sigma\bigr)\bigr].
\end{aligned}$$

Since the matrices Σ, $(I_{p^2} - C(\theta_{m,\rho_1}))$ and $(I_{p^2} - C(\widetilde{\theta}_{\rho_1}))$ correspond to covariance or precision matrices of stationary fields on the two-dimensional torus, they are symmetric block circulant. By Lemma A.1, they are jointly diagonalizable in the same orthogonal basis. In the sequel, P stands for an orthogonal matrix associated with this basis. Then the matrices $C(\theta_{m,\rho_1})$, $C(\widetilde{\theta}_{\rho_1})$ and Σ, respectively, decompose in

$$C(\theta_{m,\rho_1}) = P^* D(\theta_{m,\rho_1}) P, \qquad C(\widetilde{\theta}_{\rho_1}) = P^* D(\widetilde{\theta}_{\rho_1}) P, \qquad \Sigma = P^* D_\Sigma P,$$

where the matrices $D(\theta_{m,\rho_1})$, $D(\widetilde{\theta}_{\rho_1})$ and D_Σ are diagonal. Let the $p^2 \times n$ matrix $\mathbf{Y}$ be defined by $\mathbf{Y} := \sqrt{\Sigma^{-1}}\mathbf{X}^{\mathbf{v}}$. Clearly, the components of $\mathbf{Y}$ follow independent standard normal distributions. Gathering these new notation, we get

$$\begin{aligned}
&\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \\
&\quad = \frac{1}{p^2} \mathrm{tr}\bigl[\bigl([I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\bigr) D_\Sigma \bigl(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\bigr)\bigr].
\end{aligned} \tag{46}$$

Except $\overline{\mathbf{YY}^*}$ all the matrices in this last expression are diagonal and we may therefore commute them in the trace.

Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}'}$ be two inner products in the space of square matrices of size p^2, respectively, defined by

$$\langle A, B \rangle_{\mathcal{H}} := \frac{\operatorname{tr}(A^* \Sigma B)}{p^2} \quad \text{and} \quad \langle A, B \rangle_{\mathcal{H}'} := \frac{\operatorname{tr}(A^* D_\Sigma B)}{p^2}.$$

This first inner product is related to the loss function $l(\cdot, \cdot)$ through the identity

$$l(\theta', \theta) = \|C(\theta') - C(\theta)\|_{\mathcal{H}}^2.$$

Moreover, these two inner products clearly satisfy $\|C(\theta')\|_{\mathcal{H}} = \|D(\theta')\|_{\mathcal{H}'}$ for any $\theta' \in \Theta^+$. Gathering these new notation, we may upper bound (46) by

$$\begin{aligned}
&\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \\
&\quad \le \|[I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\|_{\mathcal{H}'} \\
&\quad\quad \times \sup_{\substack{\theta_1 \in \Theta_m, \theta_2 \in \Theta_{\widehat{m}}, \\ \|[I_{p^2} - D(\theta_1)]^2 - [I_{p^2} - D(\theta_2)]^2\|_{\mathcal{H}'} \le 1}} \langle [I_{p^2} - D(\theta_1)]^2 \\
&\qquad\qquad - [I_{p^2} - D(\theta_2)]^2, [\overline{\mathbf{YY}^*} - I_{p^2}] \rangle_{\mathcal{H}'}.
\end{aligned} \tag{47}$$

The first term in this product is easily bounded as these matrices are diagonal:

$$\begin{aligned}
&\|[I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\|_{\mathcal{H}'} \\
&\quad = \operatorname{tr}\left[\left([I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2 \right)^2 \frac{D_\Sigma}{p^2} \right]^{1/2} \\
&\quad = \operatorname{tr}\left[[D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})]^2 \frac{D_\Sigma}{p^2} [2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})]^2 \right]^{1/2} \\
&\quad \le \varphi_{\max}[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})] \|D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\|_{\mathcal{H}'}.
\end{aligned} \tag{48}$$

Since θ_{m,ρ_1} and $\widetilde{\theta}_{\rho_1}$, respectively, belong to Θ^+_{m,ρ_1} and $\Theta^+_{\widehat{m},\rho_1}$, the largest eigenvalues of the matrices $I_{p^2} - C(\theta_{m,\rho_1})$ and $I_{p^2} - C(\widetilde{\theta}_{\rho_1})$ are smaller than ρ_1. Hence, we get

$$\begin{aligned}
&\varphi_{\max}[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})] \\
&\qquad = \varphi_{\max}[I_{p^2} - C(\theta_{m,\rho_1})] + \varphi_{\max}[I_{p^2} - C(\widetilde{\theta}_{\rho_1})] \le 2\rho_1.
\end{aligned}$$

Let us turn to the second term in (47). First, we embed the set of matrices over which the supremum is taken in a ball of a vector space. For any model $m' \in \mathcal{M}$, let $U_{m'}$ be the space generated by the matrices $D(\theta')^2$ and $D(\theta')$ for $\theta' \in \Theta_{m'}$. In the sequel, we note $d_{m'^2}$ the dimension of $U_{m'}$. The space $U_{m,m'}$ is defined as the sum of U_m and $U_{m'}$ whereas d_{m^2,m'^2} stands for its dimension. Finally, we note

$\mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}$ the unit ball of $U_{m,m'}$ with respect to the inner product $\langle \ | \ \rangle_{\mathcal{H}'}$. Gathering these notation, we get

$$\sup_{\substack{R=[I-D(\theta_1)]^2-[I_{p^2}-D(\theta_2)]^2,\\ \theta_1\in\Theta_m,\theta_2\in\Theta_{\widehat{m}} \text{ and } \|R\|_{\mathcal{H}'}\le 1}} \langle R, \overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\rangle_{\mathcal{H}'} \le \sup_{R\in\mathcal{B}^{\mathcal{H}'}_{m^2,\widehat{m}^2}} \frac{1}{p^2}\operatorname{tr}[RD_\Sigma(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2})].$$

Applying the classical inequality $ab \le \delta a^2 + \delta^{-1}b^2/4$ and gathering inequalities (47) and (48) yields

$$\begin{aligned}\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\le \delta^{-1}\|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|^2_{\mathcal{H}}\\ &\quad + \rho_1^2\delta \sup_{R\in\mathcal{B}^{\mathcal{H}'}_{m^2,\widehat{m}^2}} \frac{1}{p^2}\operatorname{tr}^2[RD_\Sigma(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2})].\end{aligned} \tag{49}$$

For any model $m' \in \mathcal{M}$, we define the random variable $Z_{m'}$ as

$$Z_{m'} := \sup_{R\in\mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}} \frac{1}{p^2}\operatorname{tr}[RD_\Sigma(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2})].$$

The variables $Z_{m'}$ turn out to be suprema of Gaussian chaos of order 2. In order to bound $Z_{\widehat{m}}$, we simultaneously control the deviations of $Z_{m'}$ for any model $m' \in \mathcal{M}$ thanks to the following lemma.

LEMMA 8.3. *For any positive numbers α and ξ and any model $m' \in \mathcal{M}$,*

$$\begin{aligned}&\mathbb{P}\left(Z_{m'} \ge \sqrt{\frac{2\varphi_{\max}(\Sigma)}{n}}\{\sqrt{1+\alpha/2}\sqrt{d_{m^2,m'^2}} + \xi\}\right)\\ &\quad\le \exp\left\{-L_2\sqrt{d_{m'}}\left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2}\right) - L_1\xi\left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n}\right]\right\}.\end{aligned}$$

This result is a consequence from a general concentration inequality for suprema Gaussian chaos of order 2 stated in Proposition 8.1. Its proof is postponed to the technical Appendix [36]. Let us fix the positive numbers α and ξ. Applying Lemma 8.3 to any model $m' \in \mathcal{M}$, the event Ω'_ξ defined by

$$\Omega'_\xi = \left\{Z_{\widehat{m}} \le \sqrt{\frac{2\varphi_{\max}(\Sigma)}{n}}[\sqrt{1+\alpha/2}\sqrt{d_{m^2,\widehat{m}^2}} + \xi]\right\}$$

satisfies

$$\begin{aligned}\mathbb{P}(\Omega'^c_\xi) &\le \exp\left\{-L_1\xi\left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n}\right]\right\}\\ &\quad\times \sum_{m'\in\mathcal{M}} \exp\left\{-L_2\sqrt{d_{m'}}\left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2}\right)\right\}.\end{aligned}$$

From inequality (49), it follows that

$$\begin{aligned}\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\\ \leq \delta^{-1}\|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|^2_{\mathcal{H}} &\\ + \frac{2\delta\rho_1^2\varphi_{\max}(\Sigma)}{np^2}\{\sqrt{1+\alpha/2}\sqrt{d_{m^2,\widehat{m}^2}} + \xi\}^2,&\end{aligned}$$

conditionally to Ω'_ξ. By the triangle inequality,

$$\|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|_{\mathcal{H}} \leq \|C(\theta_{m,\rho_1}) - C(\theta)\|_{\mathcal{H}} + \|C(\widetilde{\theta}_{\rho_1}) - C(\theta)\|_{\mathcal{H}}.$$

We recall that the loss function $l(\theta', \theta)$ equals $\|C(\theta') - C(\theta)\|^2_{\mathcal{H}}$. We apply twice the inequality $(a+b)^2 \leq (1+\beta)a^2 + (1+\beta^{-1})b^2$. Setting the first β to $\sqrt{\delta} - 1$, it follows that

$$\begin{aligned}\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\\ \leq \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) &\\ + \frac{2\delta\rho_1^2\varphi_{\max}(\Sigma)}{np^2}[d_{m^2,\widehat{m}^2}(1+\beta)(1+\alpha/2) + \xi^2(1+\beta^{-1})].&\end{aligned}$$

By the definition of $U_{m,\widehat{m}}$, its dimension $d_{m^2,\widehat{m}^2}$ is bounded by $d_{m^2} + d_{\widehat{m}^2}$. Choosing $\beta = \delta - 1$ yields

$$\begin{aligned}\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\\ \leq \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) &\\ + \frac{2\delta^2\rho_1^2\varphi_{\max}(\Sigma)}{np^2}[d_{m^2}(1+\alpha/2) + d_{\widehat{m}^2}(1+\alpha/2)] &\\ + \frac{8\xi^2\varphi_{\max}(\Sigma)\delta^2}{np^2(\delta-1)}.&\end{aligned} \tag{50}$$

To conclude, we need to compare the dimension $d_{m'^2}$ of the space $U_{m'}$ with $d_{m'}$.

LEMMA 8.4. *For any model $m \in \mathcal{M}$, it holds that*

$$d_{m^2} \leq L d_m,$$

where L is a numerical constant between 4 *and* 5.48.

The proof is postponed to the technical Appendix [36]. Defining the universal constant $K_0 := 2L$, we derive from (50) that

$$\begin{aligned}\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\le \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) \\ &\quad + \frac{K_0 \delta^2 \rho_1^2 \varphi_{\max}(\Sigma)}{np^2}\left[d_m(1+\alpha/2) + d_{\widehat{m}}(1+\alpha/2) + \frac{\xi^2}{\delta-1}\right]\end{aligned}$$

with probability larger than $\mathbb{P}(\Omega'_\xi)$. The isotropic case is analogous if we replace d_m by d_m^{iso}. $\square$

8.3. *Proofs of the minimax results.* Let us first prove a minimax lower bound on hypercubes $\mathcal{C}_m(\theta', r)$. We recall that these hypercubes are introduced in Definition 6.1.

LEMMA 8.5. *Let m be a model in $\mathcal{M}_1$ that satisfies $d_m \le \sqrt{n}\,p$, and let θ' be a matrix in $\Theta_m \cap \mathcal{B}_1(0_p, 1)$. Then, for any positive number r such that $(1 - \|\theta'\|_1 - 2r d_m)$ is positive,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathrm{Co}[\mathcal{C}_m(\theta', r)]} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] \ge L\sigma^2 \left(r \wedge \frac{1 - \|\theta'\|_1}{\sqrt{np^2}}\right)^2 d_m,$$

where $\mathrm{Co}[\mathcal{C}_m(\theta', r)]$ denotes the convex hull of $C_m(\theta', r)$. Similarly, let m be a model in $\mathcal{M}_1$ such $d_m^{\mathrm{iso}} \le \sqrt{n}\,p$, and let θ' be a matrix in $\Theta_m^{\mathrm{iso}} \cap \mathcal{B}_1(0_p, 1)$. Then, for any positive number r such that $(1 - \|\theta'\|_1 - 8r d_m^{\mathrm{iso}})$ is positive,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathrm{Co}[\mathcal{C}_m^{\mathrm{iso}}(\theta', r)]} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] \ge L\sigma^2 \left(r \wedge \frac{1 - \|\theta'\|_1}{\sqrt{np^2}}\right)^2 d_m^{\mathrm{iso}}.$$

PROOF OF PROPOSITION 6.2. The first result derives from Lemma 8.5 applied to the hypercube $\mathcal{C}_m(0_p, (np^2)^{-1/2})$. We prove the second result using the same lemma with $\mathcal{C}_m[\theta', (1 - \|\theta\|_1)/(\sqrt{n}\,p)]$. $\square$

PROOF OF LEMMA 8.5. This lower bound is based on an application of Fano's approach. See [38] for a review of this method and comparisons with Le Cam's and Assouad's lemma. The proof follows three main steps: first, we upper bound the Kullback–Leibler entropy between distributions corresponding to θ_1 and θ_2 in the hypercube. Second, we find a set of points in the hypercube well separated with respect to the Hamming distance. Finally, we conclude by applying Birgé's version of Fano's lemma. More details can be found in the technical Appendix [36]. $\square$

PROOF OF PROPOSITION 6.6. First, observe that the set $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$ is included in $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(2)$. We then derive minimax lower bounds on $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$ from the lower bounds on hypercubes.

Let m_i be a model in $\mathcal{M}_1$ such that d_m is smaller than $\sqrt{n}p$. Let us look for positive numbers r such that the hypercube $[\mathcal{C}_{m_i}(0_p, r)]$ is included in the set $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$.

LEMMA 8.6. *Let m be a model in $\mathcal{M}_1$ and r be a positive number smaller than $1/(4d_m)$. For any $\theta \in \mathrm{Co}[\mathcal{C}_m(0_p, r)]$,*

$$\mathrm{var}_\theta\big(X_{[0,0]}\big) \le \sigma^2(1 + 16 d_m r^2).$$

The proof is postponed to the technical Appendix [36]. If we choose

$$r \le \frac{a_i}{16\sigma\sqrt{d_{m_i}}},$$

then $2rd_{m_i}$ is smaller than $1/8$ by assumption $(\mathbb{H}_a)$. Applying Lemma 8.6, we then derive that $\mathrm{Var}_\theta(X_{[0,0]}) \le \sigma^2 + a_i^2$. Hence, we get the upper bound $\sum_{j=1}^{i}[\mathrm{Var}(X_{[0,0]}|X_{m_{j-1}}) - \mathrm{Var}(X_{[0,0]}|X_{m_j})] \le a_i^2$ and it follows that

$$\sum_{j=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{\mathrm{Var}(X_{[0,0]}|X_{m_{k-1}}) - \mathrm{Var}(X_{[0,0]}|X_{m_j})}{a_j^2} \le 1,$$

since the sequence $(a_j)_{1 \le j \le \mathrm{Card}(\mathcal{M}_1)}$ is nonincreasing. Consequently, $\mathrm{Co}[\mathcal{C}_m(0_p, r)]$ is a subset of $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$. By Lemma 8.5, we get

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] &\ge L\sigma^2\left(\frac{a_i^2}{16\sigma^2} \wedge \frac{d_{m_i}}{np^2}\right) \\ &\ge L\left(a_i^2 \wedge \frac{\sigma^2 d_{m_i}}{np^2}\right). \end{aligned} \tag{51}$$

Considering all models $m \in \mathcal{M}_1$ such that $d_m \le \sqrt{n}p$ yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)] \ge L \sup_{i \le \mathrm{Card}(\mathcal{M}_1), d_{m_i} \le \sqrt{n}p} \left(a_i^2 \wedge \frac{\sigma^2 d_{m_i}}{np^2}\right). \tag{52}$$

If the maximal dimension $d_{m_{\mathrm{Card}(\mathcal{M}_1)}}$ is smaller than $\sqrt{n}p$, the proof is complete. In the opposite case, we need to show that the supremum (40) over all models $m \in \mathcal{M}_1$ is achieved at some model m of dimension less than $\sqrt{n}p$.

LEMMA 8.7. *For any integer $1 \le i \le \mathrm{Card}(\mathcal{M}_1) - 1$, the ratio $d_{m_{i+1}}/d_{m_i}$ is less than 2.*

The proof of Lemma 8.7 is postponed to the technical Appendix [36]. Let i' be the largest integer such that $d_{m_{i'}} \le \sqrt{n}\,p$. Since i' is smaller than $\mathrm{Card}(\mathcal{M}_1)$, we know from Lemma 8.7 that $\sqrt{n}\,p/2 \le d_{m_{i'}} \le \sqrt{n}\,p$. By assumption $(\mathbb{H}_a)$, $a_{i'}^2$ is smaller than $\sigma^2/d_{m_{i'}}$. Gathering these bounds yields

$$a_{i'}^2 \le \frac{\sigma^2}{d_{m_{i'}}} \le \frac{4d_{m_{i'}}\sigma^2}{np^2}.$$

Since the sequence $(a_i)_{1\le i\le \mathrm{Card}(\mathcal{M}_1)}$ is nonincreasing, the supremum (40) over all models in $\mathcal{M}_1$ is either achieved for some $i \le i'$ or is smaller than $4(a_{i'}^2 \wedge \sigma^2 d_{m_{i'}}/(np^2))$. $\square$

PROOF OF COROLLARY 6.3. Observe that $\mathrm{Co}[\mathcal{C}_m(0_p, 1/(4d_m)]$ is included in $\Theta_m \cap \mathcal{B}_1(0_p, 1/2)$. This last set is, itself, included in $\Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)$. Applying Lemma 8.5, we get the following minimax lower bound:

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \mathbb{E}[l(\widehat{\theta},\theta)] \ge L\sigma^2\frac{d_m}{np^2}$$

since the dimension d_m is smaller than np^2. Applying Theorem 3.1, we derive that

$$\begin{aligned}\sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \mathbb{E}[l(\widetilde{\theta}_{\rho_1},\theta)] &\le L(K)\sigma^2\rho_1^2\rho_2\frac{d_m}{np^2} \\ &\quad + L_2(K)\frac{\rho_1^2}{np^2}\sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \varphi_{\max}(\Sigma) \\ &\le L(K,\rho_1,\rho_2)\sigma^2\frac{d_m}{np^2}.\end{aligned}$$

We conclude by combining the two different bounds. $\square$

PROOF OF PROPOSITION 6.7. This result derives from the upper bound of the risk of $\widetilde{\theta}_{\rho_1}$ stated in Theorem 3.1 and the minimax lower bound stated in Proposition 6.6. For details, we refer to the technical Appendix [36]. $\square$

8.4. *Proofs of the asymptotic risk bounds.*

PROOF OF PROPOSITION 4.4. This result is closely related to Proposition 4.11 in [17]. In fact, we extend his proof to stationary fields on a torus. In the sequel, we shall only consider nonisotropic GMRFs, the isotropic case being similar. Let us fix a model m in the collection $\mathcal{M}_1$ and let us assume $(\mathbb{H}_1)$.

We define the $d_m \times p^2$ matrix χ_m^v as

$$(\chi_m^v)^* := \big([C(\Psi_{i_k,j_k})X^v], k = 1,\dots,d_m\big).$$

For any $(i,j) \in \{1,\dots,p\}^2$, the $[(i-1)p+j]$th row of χ_m^v corresponds to the list of covariates used when performing the regression of $X_{[i,j]}$ with respect to its neighbors in the model m. Contrary to the previous proofs, we need to express the $n \times p^2$ matrix $\mathbf{X}^{\mathbf{v}}$ in terms of a vector. This is why we define the vector $\mathbf{X}^{\mathbf{V}}$ of size np^2 as

$$\mathbf{X}^{\mathbf{V}}_{[p^2(j-1)+p(i_1-1)+i_2]} := \mathbf{X}^{j}_{[i_1,i_2]}$$

for any $(i_1,i_2) \in \{1,\dots,p\}^2$ and any $j \le n$. Similarly, let $\boldsymbol{\chi}_m^V$ be the $d_m \times np^2$ matrix defined as

$$\boldsymbol{\chi}^{\mathbf{V}}_{m[k,p^2(j-1)+p(i_1-1)+i_2]} := \boldsymbol{\chi}^{j}_{m[p(i_1-1)+i_2]}$$

for any $(i_1,i_2) \in \{1,\dots,p\}^2$ and any $j \le n$.

We are not able to work out directly the asymptotic risk of $\widehat{\theta}_{m,\rho_1}$. This is why we introduce a new estimator $\check{\theta}_m$ whose asymptotic distribution is easier to derive. Afterward, we shall prove that $\check{\theta}_m$ and $\widehat{\theta}_{m,\rho_1}$ have the same asymptotic distribution. Let us, respectively, define the estimators $\check{a}_m$ in $\mathbb{R}^{d_m}$ and $\check{\theta}_m$ as

$$\begin{aligned} \check{a}_m &:= ((\boldsymbol{\chi}_m^{\mathbf{V}})^* \boldsymbol{\chi}_m^{\mathbf{V}})^{-1} \boldsymbol{\chi}_m^{\mathbf{V}} \mathbf{X}^{\mathbf{V}}, \\ \check{\theta}_m &:= \sum_{k=1}^{d_m} \check{a}_{m[k]} \Psi_{i_k,j_k}, \end{aligned} \tag{53}$$

where we recall that $(\Psi_{i_1,j_1},\dots,\Psi_{i_{d_m},j_{d_m}})$ is a basis of Θ_m. Obviously, $\check{\theta}_m$ is a conditional least squares estimator since it minimizes the expression (16) of $\gamma_{n,p}(\cdot)$ over the whole space Θ_m. Consequently, $\check{\theta}_m$ coincides with $\widehat{\theta}_{m,\rho_1}$ if $\check{\theta}_m$ belongs to Θ^+_{m,ρ_1}.

For the second result, we assume that assumption $(\mathbb{H}_2)$ holds. Applying Corollary 4.2, we know that for any $(k,l) \in \Lambda$, $X_{[k,l]}$ decomposes as

$$X_{[k,l]} = \sum_{(i,j)\in m} \theta_{m,\rho_1[i,j]} X_{[k+i,l+j]} + \varepsilon_{m[k,l]}, \tag{54}$$

where $\varepsilon_{m[k,l]}$ is independent from $\{X_{[k+i,l+j]}, (i,j) \in m\}$. For the first result, the same decomposition holds since θ is assumed to belong to Θ^+_{m,ρ_1}, and θ_{m,ρ_1}, therefore, equals θ.

Let $a_m \in \mathbb{R}^{d_m}$ be the unique vector such that $\theta_{m,\rho_1} = \sum_{k=1}^{d_m} a_{m[k]} \Psi_{i_k,j_k}$. Then, the previous decomposition becomes

$$X^v = a_m^* \chi_m^v + \varepsilon_m^v.$$

Gathering this last identity with (53) yields

$$\check{a}_m - a_m = \left(\frac{1}{np^2}(\boldsymbol{\chi}_m^{\mathbf{V}})^* \boldsymbol{\chi}_m^{\mathbf{V}}\right)^{-1} \left(\frac{1}{np^2} \boldsymbol{\chi}_m^{\mathbf{V}} \boldsymbol{\varepsilon}_m^{\mathbf{V}}\right),$$

where the vector $\boldsymbol{\varepsilon}_m^{\mathbf{V}}$ of size np^2 corresponds to the n observations of the vector ε_m^v. When n goes to the infinity, $1/(np^2)(\boldsymbol{\chi}_m^{\mathbf{V}})^*\boldsymbol{\chi}_m^{\mathbf{V}}$ converges almost surely to the covariance matrix V by the law of large numbers. By definition, the variable $\varepsilon_{m[i,j]}$ is independent from the $[(i-1)p+j]$th row of $\chi_{m[i,j]}^v$. It follows that $\mathbb{E}_\theta(\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\varepsilon}^{\mathbf{V}}) = 0$. Applying again the law of large numbers we conclude that $\check{a}_m$ converges almost surely toward a_m and that $\check{\theta}_m$ converges almost surely toward θ_{m,ρ_1}. Additionally, the central limit theorem states that the random vector $1/(\sqrt{n}p)\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\varepsilon}^{\mathbf{V}}$ converges in distribution toward a zero mean Gaussian vector whose covariance matrix equals $1/p^2\operatorname{Var}_\theta(\chi_m^v\varepsilon_m^v)$. By decomposition (54), $\varepsilon_m^v = (I - C(\theta_{m,\rho_1}))X^v$ while the kth row of χ_m^v equals $[C(\Psi_{i_k,j_k})X^v]^*$. Thus for any $1 \le k,l \le d_m$,

$$\frac{1}{p^2}\operatorname{Var}_\theta(\chi_m^v\varepsilon_m^v)_{[k,l]} = \frac{1}{p^2}\operatorname{cov}_\theta\big[(X^v)^*C(\Psi_{i_k,j_k})[I - C(\theta_{m,\rho_1})]X^v,$$
$$(X^v)^*C(\Psi_{i_l,j_l})[I - C(\theta_{m,\rho_1})]X^v\big].$$

As the covariance matrix of X^v is $\sigma^2(I - C(\theta))^{-1}$, we obtain, by standard Gaussian properties,

$$\frac{1}{p^2}\operatorname{Var}_\theta(\chi_m^v\varepsilon_m^v)_{[k,l]}$$
$$= \frac{2\sigma^4}{p^2}\operatorname{cov}_\theta\big[[I - C(\theta)]^{-1}C(\Psi_{i_k,j_k})[I - C(\theta_{m,\rho_1})]$$
$$\times [I - C(\theta)]^{-1}C(\Psi_{i_l,j_l})[I - C(\theta_{m,\rho_1})]\big].$$

By Lemma A.1, all these matrices are diagonalizable in the same basis and, therefore, commute with each other. We conclude that $\frac{1}{p^2}\operatorname{Var}_\theta(\chi_m^v\varepsilon_m^v) = 2\sigma^4 W$, and

$$\sqrt{n}\,p(\check{a}_m - a_m) \to \mathcal{N}(0, V^{-1}WV^{-1}).$$

As $\widehat{\theta}_{m,\rho_1}$ belongs to Θ_{m,ρ_1}^+, there exists a unique vector $\widehat{a}_m \in \mathbb{R}^{d_m}$ such that $\widehat{\theta}_{m,\rho_1} = \sum_{k=1}^{d_m}\widehat{a}_{m[k]}\Psi_{i_k,j_k}$. The matrix θ_{m,ρ_1} belongs to the open set Θ_{m,ρ_1}^+ for the two cases of the propositions. Indeed, θ_{m,ρ_1} equals θ in the first situation. In the second situation, this is due to the fact that θ satisfies $(\mathbb{H}_2)$ and to Lemma 4.1.

Since $\check{\theta}_m$ converges almost surely to θ_{m,ρ_1}, the matrix $\check{\theta}_m$ belongs to m with probability going to one when n goes to infinity. If follows that the estimators $\check{a}_m$ and $\widehat{a}_m$ coincide with probability going to one. By Slutsky's lemma, we obtain that

$$\sqrt{n}\,p(\widehat{a}_m - a_m) \to \mathcal{N}(0, V^{-1}WV^{-1}).$$

Let us express the risk of $\widehat{\theta}_{m,\rho_1}$ with respect to the distribution of $\widehat{a}_m$:

$$l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1}) = \mathbb{E}_\theta\left[\sum_{k=1}^{d_m}(\widehat{a}_{m[k]} - a_{m[k]})\operatorname{tr}(\Psi_{i_k,j_k}X)\right]^2$$
$$= \operatorname{tr}[V(\widehat{a}_m - a_m)^*(\widehat{a}_m - a_m)].$$

By Portmanteau's lemma, $np^2 l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})$ converges in distribution toward a random variable whose expectation is $\mathrm{tr}(WV^{-1})$. In order to conclude, it remains to prove that the sequence $[np^2 l(\widehat{\theta}_{m,\rho_1}, \theta)]_{n\geq 1}$ is asymptotically uniformly integrable.

Let us consider a model selection procedure with the collection $\mathcal{M} = \{m\}$ and a penalty term satisfying the assumptions of Theorem 3.1. Arguing as in the proof of this theorem, we derive from identity (44) the following property. For any $\xi > 0$, with probability larger than $1 - L_1 \exp[-L_2\xi]$,

$$np^2 l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1}) \leq L_3 d_m \varphi_{\max}(\Sigma) + L_4 \xi^2 \varphi_{\max}(\Sigma).$$

This clearly implies that the sequence $[np^2 l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]_{n\geq 1}$ is asymptotically uniformly integrable and the first part of the result follows.

For the first result of the proposition, we have stated that θ equals Θ_m. As a consequence,

$$\lim_{n\to+\infty} \mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)] = 2\sigma^4 \,\mathrm{tr}[WV^{-1}].$$

Also, the term $W_{[k,l]}$ here equals $\mathrm{tr}[C(\Psi_{i_k,j_k})C(\Psi_{i_l,j_l})]$. This last quantity is zero if $k \neq l$ and equals $\|C(\Psi_{i_k,j_k})\|_F^2$ if $k = l$. □

PROOF OF PROPOSITION 4.7. As θ belongs to $\Theta^+ \cap \mathcal{B}_1(0_p, \eta)$, the largest eigenvalue of Σ is smaller than $\sigma^2/(1-\eta)$. Applying Theorem 3.1, we get

$$\begin{aligned}\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)] &\leq L(K) \inf_{m\in\mathcal{M}} \left[l(\theta_{m,\rho_1}, \theta) + K \frac{\sigma^2}{np^2(1-\eta)} \right] \\ &\leq L(K,\eta) \inf_{m\in\mathcal{M}} \left[l(\theta_{m,\rho_1}, \theta) + K \frac{\sigma^2}{np^2}(1-\eta)^3 \right].\end{aligned}$$

Gathering this bound with the result of Corollary 4.6 enables us to conclude. □

APPENDIX

LEMMA A.1. *There exists an orthogonal matrix P which simultaneously diagonalizes every $p^2 \times p^2$ symmetric block circulant matrices with $p \times p$ blocks. Conversely, if θ is a square matrix of size p which satisfies (3), then the matrix $D(\theta) = PC(\theta)P^*$ is diagonal and satisfies*

$$D(\theta)_{[(i-1)p+j,(i-1)p+j]} = \sum_{k=1}^{p}\sum_{l=1}^{p} \theta_{[k,l]} \cos\big(2\pi(ki/p + lj/p)\big) \tag{55}$$

for any $1 \leq i, j \leq p$.

This lemma is proved in [29], Section 2.6.2 when is P a unitary matrix. A slight modification of their proof allows to show that P is orthogonal in our case. The difference comes from the fact that contrary to Rue and Held we also assume that $C(\theta)$ is symmetric.

This lemma states that all symmetric block circulant matrices are simultaneously diagonalizable. Moreover, expression (55) explicitly provides the eigenvalues of the $C(\theta)$ as the two-dimensional discrete Fourier transform of the $p \times p$ matrix θ.

Acknowledgments. I am grateful to Pascal Massart for many fruitful discussions. I also thank the referees and the associate editor for their suggestions that led to an improvement of the manuscript.

REFERENCES

[1] AYKROYD, R. (1998). Bayesian estimation for homogeneous and inhomogeneous Gaussian random fields. *IEEE Trans. Pattern Anal. Machine Intell.* **20** 533–539.
[2] BESAG, J. E. (1975). Statistical analysis of non-lattice data. *Statistica* **24** 179–195.
[3] BESAG, J. E. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64** 616–618. MR0494640
[4] BESAG, J. E. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811
[5] BESAG, J. E. and MORAN, P. A. P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* **62** 555–562. MR0391451
[6] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268. MR1848946
[7] BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064
[8] BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200
[9] BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed. Springer, New York. MR1093459
[10] CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641
[11] CRESSIE, N. A. C. and VERZELEN, N. (2008). Conditional-mean least-squares of Gaussian Markov random fields to Gaussian fields. *Comput. Statist. Data Anal.* **52** 2794–2807. MR2419542
[12] CROUSE, M., NOWAK, R. and BARANIUK, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46** 886–902. MR1665651
[13] DASS, S. C. and NAIR, V. N. (2003). Edge detection, spatial smoothing, and image reconstruction with partially observed multivariate data. *J. Amer. Statist. Assoc.* **98** 77–89. MR1977200
[14] EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer, New York. MR1880319
[15] GRAY, R. (2006). *Toeplitz and Circulant Matrices: A Review*, rev. ed. Now Publishers, Norwell, MA.
[16] GUYON, X. (1987). Estimation d'un champ par pseudo-vraisemblance conditionnelle: Étude asymptotique et application au cas Markovien. In *Spatial processes and spatial time series analysis (Brussels, 1985). Travaux Rech.* **11** 15–62. Publ. Fac. Univ. Saint-Louis, Brussels. MR0947996
[17] GUYON, X. (1995). *Random Fields on a Network*. Springer, New York. MR1344683
[18] GUYON, X. and YAO, J. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Multivariate Anal.* **70** 221–249. MR1711522

[19] HALL, P., FISHER, N. and HOFFMANN, B. (1994). On the nonparametric estimation of covariance functions. *Ann. Statist.* **22** 2115–2134. MR1329185
[20] HURVICH, C. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020
[21] IM, H., STEIN, M. and ZHU, Z. (2007). Semiparametric estimation of spectral density with irregular observations. *J. Amer. Statist. Assoc.* **102** 726–735. MR2381049
[22] KASHYAP, R. and CHELLAPA, R. (1984). Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Trans. Inform. Theory* **29** 60–72. MR0781270
[23] LAKSHMANAN, S. and DERIN, H. (1993). Valid parameter space for 2-D Gaussian Markov random fields. *IEEE Trans. Inform. Theory* **39** 703–709. MR1224361
[24] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991
[25] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879
[26] MCQUARRIE, A. D. R. and TSAI, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific, River Edge, NJ. MR1641582
[27] PORTILLA, J., STRELA, V., WAINWRIGHT, M. J. and SIMONCELLI, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.* **12** 1338–1351. MR2026777
[28] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391
[29] RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall/CRC, London. MR2130347
[30] RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 319–392.
[31] RUE, H. and TJELMELAND, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.* **29** 31–49. MR1894379
[32] SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164. MR0557560
[33] SONG, H.-R., FUENTES, M. and GHOSH, S. (2008). A comparative study of Gaussian geostatistical models and Gaussian Markov random field models. *J. Multivariate Anal.* **99** 1681–1697.
[34] STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York. MR1697409
[35] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006
[36] VERZELEN, N. (2009). Technical Appendix to "Adaptive estimation of stationary Gaussian fields." Available at arXiv:0908.4586.
[37] VERZELEN, N. (2010). Data-driven neighborhood selection of a Gaussian field. *Comput. Statist. Data Anal.* To appear.
[38] YU, B. (1997). Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

INRA AND SUPAGRO
UMR 729 MISTEA
BÂTIMENT 29
2, PLACE PIERRE VIALA
F-34060 MONTPELLIER
FRANCE
E-MAIL: nicolas.verzelen@supagro.inra.fr