Adaptive estimation of stationary Gaussian fields

Nicolas Verzelen

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Adaptive estimation of stationary Gaussian fields

Nicolas Verzelen

N° 6797

Janvier 2009

Thème COG

apport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6797--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Adaptive estimation of stationary Gaussian fields

Nicolas Verzelen * †

Thème COG — Systèmes cognitifs
Équipes-Projets Select

Rapport de recherche n° 6797 — Janvier 2009 — 58 pages

Abstract: We study the non-parametric 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.

Key-words: Gaussian field, Gaussian Markov random field, model selection, pseudolikelihood, oracle inequalities, Minimax rate of estimation.

* Laboratoire de Mathématiques UMR 8628, Université Paris-Sud, 91405 Osay
† INRIA Futurs, Projet SELECT, Université Paris-Sud, 91405 Osay

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Estimation adaptative de champs gaussiens stationnaires

Résumé : Nous étudions l'estimation non-paramétrique d'un champ gaussien stationnaire X observé sur un réseau régulier. Dans le cadre des séries temporelles, certaines procédures comme AIC réalisent une sélection de modèle optimale parmi les modèles autorégressifs. Cependant, il n'existe aucun résultat analogue d'adaptation pour des champs spatiaux. En considérant des collections de champs de Markov gaussiens comme des ensembles d'approximation de la distribution de X, nous introduisons une nouvelle méthode de sélection de modèle pour des champs spatiaux. Pour tout voisinage m dans une collection $\mathcal{M}$ donnée, cette procédure estime la covariance de X par un champ de Markov de voisinage m. Puis, elle sélectionne un voisinage $\widehat{m}$ grâce à une technique de pénalisation. L'estimateur ainsi défini satisfait une inégalité oracle non-asymptotique. Si X est un champ de Markov gaussien, la procédure est minimax adaptative à la taille de son voisinage. Plus généralement, nous prouvons que la procédure s'adapte à la vitesse d'approximation de la distribution de X par des champs de Markov gaussiens de voisinage croissant.

Mots-clés : Champ gaussien, champ de Markov gaussien, sélection de modèle, pseudo-vraisemblance, inégalités oracles, vitesse minimax d'estimation.

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 popular approach amounts to first computing an empirical variogram and then fitting a suitable parametric variogram model such as the exponential or Matérn model. It is beyond the scope of this paper to do an exhaustive review of these methods and we refer to [Cre93] Ch.2 for more details. Some procedures also apply to non-regular 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 [Cre93] Sect.2.6.4 for a discussion. Alternatively, Rosenblatt [Ros85] Ch.5 has developed a non-parametric estimator of the spectral density of the field X. This procedure is shown to be universally consistent, but it fails to achieve the optimal rate of convergence when the true distribution belongs to one parametric model.

On the other hand, a second 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 satisfies the local Markov property 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. The field X is said to be a GMRF with respect to the graph $\mathcal{G}$ if it fulfills the local Markov property with respect to $\mathcal{G}$. 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. See [Lau96] and [Edw00] 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 here 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 $(0,0)$). Let us mention the idea underlying our approach: using the same data, we select a *suitable* neighborhood and estimate the distribution of X in the space of stationary GMRFs with respect to this neighborhood.

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 [BM75, Bes77, Guy87]). If the field X is assumed to be a GMRF with respect to a *known* neighborhood in all these works, the issue of neighborhood selection has been less studied. Besag and Kooperberg [BK95], Rue and Tjelmeland [RT02], Song *et al.* [SFG08], and Cressie and Verzelen [CV08] 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 [GY99] necessary conditions and sufficient conditions for a model selection procedure to choose asymptotically the true neighborhood of a GMRF with probability one. Our point of

view is slightly different: we do not assume that the field X is a GMRF with respect to a sparse neighborhood and do not aim at estimating the true neighborhood, we rather want to select a neighborhood that allows to estimate *well* the distribution of X.

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 [Guy95] Sect. 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 [Shi80] and Hurvich and Tsai [HT89]). We refer to Brockwell and Davis [BD91] and McQuarrie and Tsai [MT98] for detailed discussions. However, one cannot readily extend these results to a spatial setting because of computational and theoretical difficulties.

1.1 Conditional regression

Let us now precise the notations 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 $X = (X_{[i,j]})_{1 \leq i,j \leq p}$ indexed by the nodes of Λ. Besides, X^v refers to the vectorialized version of X with the convention $X_{[i,j]} = {X^v}_{[(i-1)\times p+j]}$ for any $1 \leq i, j \leq p$. Using this new notation amounts to "forgetting" the spatial structure of X and allows 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 non singular but unknown. Besides, we suppose that the field X is stationary on the torus Λ. More precisely, for any $r > 0$, any $(i, j) \in \{1, \ldots, p\}^2$, and any $(k_1, l_1), \ldots, (k_r, l_r) \in \{1, \ldots, p\}^{2r}$, it holds that

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

We observe $n \geq 1$ 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 $1 \leq i \leq n$, the $p \times p$ matrix $\mathbf{X}_i$ stands for the i-th observation of the field X. All these notations are recalled in Table 1 in Section 1.4. 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 $(k_1, l_1), \ldots, (k_r, l_r) \in \{1, \ldots, p\}^{2r}$,

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

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 [Lau96] App.C), there exists a unique $p \times p$ matrix θ such that $\theta_{[0,0]} = 0$ and

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

where the random variable $\epsilon_{[0,0]}$ 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:

$$\Sigma = \sigma^2 \left[I_{p^2} - C(\theta) \right]^{-1} , \tag{2}$$

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 $1 \leq i_1, i_2, j_1, j_2 \leq p$. 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 [RH05] Sect.2.6 or the book of Gray [Gra06] 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*

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

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 appendix. 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 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

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

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 $\Theta^{+,\text{iso}}$ for isotropic Gaussian fields.

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

1.2 Model selection

The issue of covariance estimation may be reformulated as a problem of conditional regression defined in Equation (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 non-independent 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 Equation (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. We do not want to perform a restrictive assumption on the true distribution. In general, we aim at adapting 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 Equation (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, $\gamma_{n,p}(.)$ stands for the CLS empirical contrast that we shall define in Section 2. We recall that it is closely connected to the pseudolikelihood. We select a model $\widehat{m}$ by minimizing the criterion

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

where pen(.) denotes a positive function defined on $\mathcal{M}$. In this paper, we prove that under a suitable choice of the penalty function pen(.), 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 $l(.,.)$ 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

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

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

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

where $\mathbb{E}_\theta(.)$ 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 [HT89], or in spatial statistics [SFG08]. Moreover, the loss function $l(\widehat{\theta}, \theta)$ is also connected to the notion of kriging error. The kriging predictor 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 Equation (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 refer to Cressie [Cre93] Ch.2 for a comprehensive introduction on kriging and related notions.

We shall also assess our results thanks to the Frobenius distance noted $\|.\|_F$ and defined by $\|A\|_F^2 := \sum_{1\leq i,j\leq p} A_{[i,j]}{}^2$. 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)

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

Our aim is then to define a suitable penalty function pen(.) 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 pen(.), so that the selected estimator satisfies a risk bound of the form

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

where $\varphi_{\max}(\Sigma)$ is the largest eigenvalue of Σ. Contrary to most results in a spatial setting, this upper bound on the risk is nonasymptotic and holds in a general setting. The term $\varphi_{\max}(\Sigma)\frac{\mathrm{Card}(m)}{np^2}$ 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 $\varphi_{\max}(\Sigma)\frac{\mathrm{Card}(m)}{np^2}$. 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 $l(.,.)$ and $\|.\|_F^2$. 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 [Shi80]. 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 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 made in a subsequent paper [Ver09]. All the proofs are postponed to Section 8 and to the appendix.

1.4 Some notations

Throughout this paper, $L, L_1, L_2, \ldots$ denote constants that may vary from line to line. The notation $L(.)$ 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

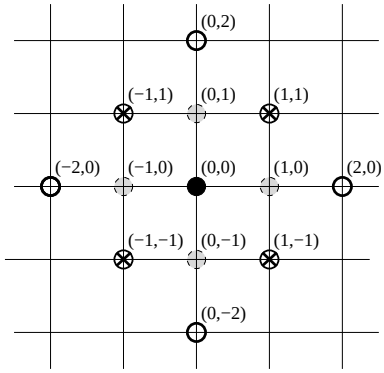

Figure 1: *Examples of models. The four gray nodes refer to* m_1*. The model* m_2 *also contains the nodes with a cross whereas* m_3 *contains all the nodes except* $(0,0)$*.*

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 notations involving X.

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 $p \times p$	i-th observation of the field X

Table 1: Notations for the random field and the data.

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*

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

The collection $\mathcal{M}_1$ is totally ordered with respect to the inclusion and we therefore order our models $m_0 \subset m_1 \subset \ldots \subset m_i \ldots$. 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 aim at estimating the positive matrix $(I_{p^2} - C(\theta))$, we shall consider the convex subsets of Θ_m^+ and $\Theta_m^{+,\text{iso}}$ that correspond to non-negative precision matrices.

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

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 $\Theta_{m,\rho_1}^{+,\text{iso}}$ for any $\rho_1 \geq 2$.

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

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

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

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

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 $\Psi_{i,j}^{\text{iso}}$ by

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

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

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

In a nutshell, $\gamma_{n,p}(\theta')$ is a least squares criterion that allows 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 $\overline{\mathbf{X}^{\mathbf{v}} \mathbf{X}^{\mathbf{v}*}}$:

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

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 $\widehat{\theta}^{\text{iso}}_{m,\rho_1}$ by minimizing the criterion $\gamma_{n,p}(.)$ as follows

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

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

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

The proof is postponed to the appendix. 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 $\gamma_{n,p}(.)$ over a set Θ^+_{m,ρ_1} is equivalent to minimizing the product of the conditional likelihoods $\left(X_{[i,j]}|X_{-\{i,j\}}\right)$, called *Conditional Pseudo-Likelihood* (CPL):

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

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 [Bes75] who call them pseudolikelihood estimators since they minimize the CPL.

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

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

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 in (Eq.10) of [CV08] 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 $\gamma(.)$ is minimized over Θ^+ at the point θ and that $\gamma(\theta) = \text{Var}_\theta\left(X_{[0,0]}\left|X_{-\{0,0\}}\right.\right) = \sigma^2$. 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 $\theta^{\text{iso}}_{m,\rho_1}$ as the best approximation of θ in $\overline{\Theta^+_{m,\rho_1}}$ and $\overline{\Theta^{+,\text{iso}}_{m,\rho_1}}$.

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

Since $\gamma(.)$ is strictly convex on Θ^+, the matrices θ_{m,ρ_1} and $\theta^{\text{iso}}_{m,\rho_1}$ are uniquely defined. By its definition (7), one may interpret $l(.,.)$ 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. $\overline{\Theta^{+,\text{iso}}_{m,\rho_1}}$) with respect to $l(.,.)$

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

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

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.

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

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. $\varphi_{\min}(\Sigma) \leq \frac{\sigma^2}{\rho_1}$. 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 $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ that we call a penalty, we select a model as follows

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

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 $m \in \mathcal{M}$,*

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

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

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

A similar bound holds if one replaces $\widetilde{\theta}_{\rho_1}$ by $\widetilde{\theta}^{\mathrm{iso}}_{\rho_1}$, Θ^+ by $\Theta^{+,iso}$, 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 and proved 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

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

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.

Comments:

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 $\text{pen}(m) = K\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$, we get

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

   We shall prove in Section 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 the 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.

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

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

   This is why we want to compare the sum $l(\theta_{m,\rho_1},\theta)+\text{pen}(m)$ 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 $\varphi_{\max}(\Sigma)$.

3. Condition (23) gives a lower bound on the penalty function pen(.) 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,*

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

A similar result holds for isotropic GMRFs.

Proof of Corollary 3.2. This is a consequence of Theorem 3.1. By definition (7) of the loss function $l(.,.)$, the two following bounds hold

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

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 $\|C(\theta_1) - C(\theta_2)\|_F^2 = p^2\|\theta_1 - \theta_2\|_F^2$:

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

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 the projection θ_{m,ρ_1} belongs to Θ^+_m. This is why we introduce the condition $(\mathbb{H}_2)$.

$$(\mathbb{H}_2): \qquad \theta \in \mathcal{B}_1(0_p, 1) \qquad \Longleftrightarrow \qquad \|\theta\|_1 < 1. \tag{27}$$

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 $\gamma(.)$ over Θ_m is achieved in $\Theta^+_{m,2}$. This implies that*

$$\theta_{m,\rho_1} = \arg\min_{\theta'\in\Theta_m}\gamma(\theta') \qquad \text{and} \qquad \gamma(\theta_{m,\rho_1}) = Var_\theta\left(X_{[0,0]}|X_m\right)\ .$$

Besides, $\|\theta_{m,\rho_1}\|_1 \leq \|\theta\|_1$. The same results holds for θ^{iso}_{m,ρ_1} if θ in $\Theta^{+,\mathrm{iso}}$.

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*

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

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 $\theta^{\mathrm{iso}}_{m,\rho_1}$ is uniquely defined by the equation

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

and $\theta^{\mathrm{iso}}_{m,\rho_1}[i,j] = 0$ 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 (see Cressie [Cre93] Ch.3 for an introduction). 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 $l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1},\theta)=l(\theta^{\mathrm{iso}}_{m,\rho_1},\theta)+l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1},\theta_{m,\rho_1})$.*

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 [Guy95] Sect.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 notations. For any model m in the collection $\mathcal{M}_1\setminus\{\emptyset\}$, 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 $X_{[0,0]}$

$$\chi_{m[0,0]}{}^*:=\left[tr\left(\Psi_{i_1,j_1}X^v\right),\dots,tr\left(\Psi_{i_{d_m},j_{d_m}}X^v\right)\right]\ .$$

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

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

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 $\mathcal{M}_1\setminus\{\emptyset\}$ 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*

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

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

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

Both results still hold for the estimator $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ if θ belongs to $\Theta^{+,\mathrm{iso}}$ and if one replace 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.

The 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 the $(X_{[i,j]})_{(i,j)\in\Lambda}$.

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 $\widehat{\theta}^{\text{iso}}_{m,\rho_1}$ satisfy*

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

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 non zero, 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*

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

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

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

Again, analogous lower bounds hold for $\widehat{\theta}^{\text{iso}}_{m,\rho_1}$ when θ belongs to $\Theta^{\text{iso},+}$. This corollary states that asymptotically with respect to n the variance term of $\widehat{\theta}_{m,\rho_1}$ is larger than the order $\frac{d_m}{np^2}$. 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 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 the corollary (30) states that from an asymptotic point of view the upper bound (25) is optimal. By Theorem 3.1, if we choose $\text{pen}(m) = K\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$, then it holds that

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

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

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

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 non zero element of $\Theta^{\text{iso}}_{m_1}$, then the asymptotic risk of $\widehat{\theta}^{\text{iso}}_{m_1,\rho_1}$ simplifies as*

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

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

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

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 $\mathbb{E}_\theta[l(\widehat{\theta}^{iso}_{m_1,\rho_1}, \theta)]$ behaves like $\sigma^2(1-\|\theta\|_1)d^{\text{iso}}_{m_1}/(np^2)$ whereas the penalty $\text{pen}(m_1)$ given in Theorem 3.1 has to be larger than $\sigma^2 d^{\text{iso}}_{m_1}/[np^2(1-\|\theta\|_1)]$. Hence, the variance term and the penalty pen(.) 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*

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

Then, the variance term is asymptotically lower bounded as follows

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

Consequently, this variance is of order $\sigma^2 \frac{d_m^{\text{iso}}}{np^2(1-\|\theta\|_1)} = \varphi_{\max}(\Sigma)\frac{d_m^{\text{iso}}}{np^2}$ when $\|\theta\|_1$ goes to one. The penalty pen(m) introduced in Proposition 4.7 is therefore a sharp upper bound of the variance terms.

On the one hand, we take a penalty pen(m) larger than $\sigma^2 \frac{d_m}{np^2(1-\|\theta\|_1)}$, whereas in some cases the variance of $\widehat{\theta}_{m,\rho_1}$ is of order $\sigma^2(1-\|\theta\|_1)\frac{d_m}{np^2}$. The 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 $\gamma_{n,p}(.)$ 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

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

The quantities $\rho(m)$ and $\rho^{iso}(m)$ are finite since Θ^+_m is bounded. If one takes ρ_1 larger than $\rho(m)$ (resp. $\rho^{\text{iso}}(m)$), then the set Θ^+_{m,ρ_1} (resp. $\Theta^{+,\text{iso}}_{m,\rho_1}$) is exactly Θ^+_m (resp. $\Theta^{+,\text{iso}}_m$). We illustrate in Table 2 that $\rho(m)$ and $\rho^{\text{iso}}(m)$ are small, when the model m is small. Consequently, choosing a moderate value for ρ_1 is not really restrictive for small models. However, when the size of the 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.

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

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

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 [RH05] Sect.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 [Guy95]).

If one looks closely at our proofs involving Assumptions ($\mathbb{H}_2$), one realizes that this assumptions 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

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

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 [BM01] 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

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

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 $\mathcal{M}_1 \setminus \emptyset$. We consider $\left(\Psi_{i_1,j_1}, \ldots, \Psi_{i_{d_m},j_{d_m}}\right)$ 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 $\left(\Psi^{\mathrm{iso}}_{i_1,j_1}, \ldots, \Psi^{\mathrm{iso}}_{i_{d_m},j_{d_m}}\right)$.

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

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

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

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

where $\mathrm{Co}\left[\mathcal{C}_m(\theta', r)\right]$ 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 $\sigma^2 \frac{d_m}{np^2}$ (up to some numerical constant). This rate $\sigma^2 \frac{d_m}{np^2}$ 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 Proposition 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 $pen(m) = K\sigma^2 \rho_1^2 \rho_2 \frac{d_m}{np^2}$. For any non empty model m,*

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

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 and 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 the smallest and 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,*

$$\sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \mathbb{E}_\theta\left[\|C(\widetilde{\theta}_{\rho_1})-C(\theta)\|_F^2\right] \leq L\left(K,\rho_1,\rho_2\right)\inf_{\widehat{\theta}}\sup_{\theta\in\Theta^+_{m,\rho_1}\cap\mathcal{U}(\rho_2)} \mathbb{E}\left[\|C(\widehat{\theta})-C(\theta)\|_F^2\right] \ . \quad (38)$$

Proof of Corollary 6.4. As in the proof of Corollary 3.2, we observe that

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

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 $(a_j)_{1\leq j\leq Card(\mathcal{M}_1)}$ be a non-increasing sequence of positive numbers. Then, $\theta\in\Theta^+$ belongs to the* pseudo-ellipsoid $\mathcal{E}(a)$ *if and only if*

$$\sum_{i=1}^{Card(\mathcal{M}_1)} \frac{\mathrm{var}_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_{i-1})}\right)-\mathrm{var}_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_i)}\right)}{a_i^2} \quad \leq \quad 1 \ . \quad (39)$$

Condition (39) 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, $\mathrm{Var}_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_i)}\right)$ is the sum of $l(\theta_{m_i},\theta)$ and σ^2 and Condition (39) is equivalent to

$$\sum_{i=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{l\left(\theta_{m_{i-1}},\theta\right)-l(\theta_{m_i},\theta)}{a_i^2} \leq 1 \ . \quad (40)$$

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 [Mas07] Sect.4.3).

To prove adaptivity, we shall need the equivalence between Conditions (39) and (40). This equivalence holds if $\mathrm{Var}_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_i)}\right)$ 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 $\mathcal{E}(a)$

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

It essentially means that the sequence (a_i) converges fast enough towards 0. For instance, all the sequences $a_i=\sigma(d_{m_i})^{-s}$ with $s\geq 1/2$ 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*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]\geq L\sup_{1\leq i\leq Card(\mathcal{M}_1)}\left(a_i^2\wedge\sigma^2\frac{d_{m_i}}{np^2}\right)\ . \tag{41}$$

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

$$\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right]\leq L(K,\rho_1,\rho_2)\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]\ . \tag{42}$$

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 $s\geq 1/2$. We note $\mathcal{E}'(s)$ 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 Proposition 6.7. By (42), 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 $a_1^2\geq\frac{1}{np^2}$. 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, i.e. that the minimax rates of estimation is not smaller than $\sigma^2/(np^2)$. We have explained earlier that we restricts ourselves to parameters θ in $\mathcal{B}_1(0_p,1)$ only because this enforces the equivalence between (39) and (40). 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*

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

Under the same assumptions as Corollary 6.7,

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

Proof of Corollary 6.8. As in the proof of Corollary 3.2, we observe that

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

$$\|C(\theta_1)-C(\theta_2)\|_F\leq p^2[\varphi_{\min}(\Sigma)]^{-1}l(\theta_1,\theta_2)\leq p^2\frac{\varphi_{\max}[I_{p^2}-C(\theta)]}{\sigma^2}l(\theta_1,\theta_2)\leq\frac{\rho_2p^2}{\sigma^2}l(\theta_1,\theta_2)\ ,$$

if $\theta\in\mathcal{B}_1(0_p,1)\cap\mathcal{B}_{\mathrm{op}}(\rho_2)$. We conclude by applying Proposition 6.6 and Corollary 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 [RT02]) 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 [Ver09] Sect.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 $\gamma_{n,p}(.)$ defined in (16) is a quadratic function of θ. Consequently, CLS estimators are still computationally amenable. We extend our model selection to non-toroidal lattices in [Ver09].

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 ([Guy95] Sect.4.3). It is generally accepted (see for instance Cressie [Cre93] Sect. 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 non-asymptotically 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 [KC84] 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 $l(.,.)$. 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 (42), 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 Proposition 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 [Guy95] Sect.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 done in [RBLZ08] in the non-stationary 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 the 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 [Ver09], 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 define in [Ver09] a data-driven method based on the *slope heuristics* of Birgé and Massart [BM07] for calibrating the penalty. We also provide numerical evidence of its performances on simulated data.

We 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 non asymptotic 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 [LD93] 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 [Ver09], we describe such an extension and we provide numerical evidence of its performances.

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*

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

Then

$$\mathbb{P}(Z \geq \mathbb{E}(Z) + t) \leq \exp\left[-\left(\frac{t^2}{L_1 \mathbb{E}(W)} \bigwedge \frac{t}{L_2 B}\right)\right], \tag{43}$$

where the quantities B and W are such that

$$\begin{aligned} B &:= \frac{2}{n} \sup_{R \in F} \varphi_{max}(R) \\ W &:= \frac{4}{n} \sup_{R \in F} tr(R\overline{YY^*}R') \ . \end{aligned}$$

Proof of Proposition 8.1. 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 [Tal96] has first given in Theorem 1.2 a concentration inequality for such suprema of Rademacher chaos. Boucheron *et al.* [BBLM05] have recovered the upper bound applying a new methodology based on the entropy method. We shall adapt their proof to consider non-necessarily homogeneous chaos of order 2.

First, we recall the notations introduced in [BBLM05]. Let N be a positive integer. Then, $\mathcal{I}_N$ stands for the family of subsets of $\{1, \ldots, N\}$ of size less than 2. Let $\mathcal{T}$ be a set of vectors indexed by $\mathcal{I}_N$. In the sequel, $\mathcal{T}$ is assumed to be a compact subset of $\mathbb{R}^{(N(N+1)/2)+1}$. The following lemma states a slightly modified version of the upper bound in remark 7 in [BBLM05].

Lemma 8.2. *Let T be a suprema of Rademacher chaos indexed by $\mathcal{I}_N$ of the form*

$$T := \sup_{t \in \mathcal{T}} \left| \sum_{\{i,j\}} U_i U_j t_{\{i,j\}} + \sum_{i=1}^{N} t_{\{i\}} + t_{\varnothing} \right| \ ,$$

where $U_1, \ldots, U_N$ are independent Rademacher random variables. Then for any $x > 0$,

$$\mathbb{P}\left\{T \geq \mathbb{E}[T] + x\right\} \leq 4 \exp\left(-\frac{x^2}{L_1 \mathbb{E}[D]^2} \wedge \frac{x}{L_2 E}\right) \ , \tag{44}$$

where D and E are defined by:

$$\begin{aligned} D &:= \sup_{t \in \mathcal{T}} \sup_{\alpha : \|\alpha\|_2 \leq 1} \left| \sum_{i=1}^{N} U_i \sum_{j \neq i} \alpha_j t_{\{i,j\}} \right| \ , \\ E &:= \sup_{t \in \mathcal{T}} \sup_{\alpha^{(1)}, \alpha^{(2)}, \|\alpha^{(1)}\|_2 \leq 1 \ \|\alpha^{(2)}\| \leq 1} \left| \sum_{i=1}^{N} \sum_{j \neq i} t_{\{i,j\}} \alpha_i^{(1)} \alpha_j^{(2)} \right| \ . \end{aligned}$$

Contrary to the original result of [BBLM05], the chaos are not assumed to be homogeneous. Besides, the $t_{\{i\}}$ are redundant with $t_{\varnothing}$. In fact, we introduced this family in order to emphasize the connection with Gaussian chaos in the next result.

A suitable application of the central limit theorem enables to obtain a corresponding bound for Gaussian chaos of order 2.

Lemma 8.3. *Let T be a supremum of Gaussian chaos of order 2.*

$$T := \sup_{t \in \mathcal{T}} \left| \sum_{\{i,j\}} t_{\{i,j\}} Y_i Y_j + \sum_i t_i Y_i^2 + t_{\varnothing} \right| \ , \tag{45}$$

where $Y_1, \dots, Y_N$ are independent standard Gaussian random variable. Then, for any $x > 0$,

$$\mathbb{P}\left\{T \geq \mathbb{E}[T] + x\right\} \leq \exp\left(-\frac{x^2}{\mathbb{E}[D]^2 L_1} \wedge \frac{x}{EL_2}\right) , \tag{46}$$

where

$$\begin{aligned} D &:= \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^N \|\alpha\|_2\leq 1} \sum_{i,j} Y_i(1+\delta_{i,j})\alpha_j t_{\{i,j\}} , \\ E &:= \sup_{t\in\mathcal{T}} \sup_{\alpha_1,\ \|\alpha_1\|_2\leq 1} \sup_{\alpha_2,\ \|\alpha_2\|_2\leq 1} \sum_{i,j} \alpha_{1,i}\alpha_{2,j} t_{\{i,j\}}(1+\delta_{i,j}) . \end{aligned}$$

The proof of this Lemma is postponed to the end of this section. To conclude, we derive the result of Proposition 8.1 from this last lemma. For any matrix $R \in F$, we define the vector $t^R \in \mathbb{R}^{nr(nr+1)/2+1}$ indexed by $\mathcal{I}_{nr}$ as follows

$$t^R_{\{(i,k),(j,l)\}} := \delta_{k,l}(2-\delta_{i,j})\frac{R_{[i,j]}}{n}, \quad t^R_{\{(i,k)\}} := \frac{R_{[i,i]}}{n}, \quad \text{and } t^R_{\varnothing} := -tr(R) ,$$

where $\delta_{i,j}$ is the indicator function of $i = j$. In order to apply Lemma 8.3 with $N = nr$ and $\mathcal{T} = \left\{t^R | R \in F\right\}$, we have to work out the quantities D and E.

$$\begin{aligned} D &= \sup_{t^R\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nr},\ \|\alpha\|_2\leq 1} \left\{\sum_{i=1}^r\sum_{k=1}^n Y_{[i,k]} \sum_{j=1}^r\sum_{l=1}^n t^{R,k,l}_{ij}(1+\delta_{i,j}\delta_{k,l})\alpha^l_j\right\} \\ &= \sup_{R\in F} \sup_{\alpha\in\mathbb{R}^{nr},\ \|\alpha\|_2\leq 1} 2\left\{\sum_{i=1}^r\sum_{k=1}^n Y_{[i,k]} \sum_{j=1}^r \frac{R_{[i,j]}\alpha^k_j}{n}\right\} \\ &= \sup_{R\in F} \sup_{\alpha\in\mathbb{R}^{nr},\ \|\alpha\|_2\leq 1} \frac{2}{n}\left\{\sum_{k=1}^n\sum_{j=1}^r \alpha^k_j\left(\sum_{i=1}^r Y_{[i,k]}R_{[i,j]}\right)\right\} . \end{aligned}$$

Applying Cauchy-Schwarz identity yields

$$\begin{aligned} D^2 &= \frac{4}{n^2}\sup_{R\in F}\left\{\sum_{k=1}^n\sum_{j=1}^r\left(\sum_{i=1}^r Y_{[i,k]}R_{[i,j]}\right)^2\right\} \\ &= \frac{4}{n}\sup_{R\in F} tr(R\overline{YY^*}R^*) . \end{aligned} \tag{47}$$

Let us now turn the constant E

$$\begin{aligned} E &= \sup_{t^R\in\mathcal{T}} \sup_{\substack{\alpha_1,\alpha_2\in\mathbb{R}^{nr} \\ \|\alpha_1\|_2\leq 1, \|\alpha_2\|_2\leq 1}} \sum_{1\leq i,j\leq r}\sum_{1\leq k,l\leq n}(1+\delta_{ij}\delta_{k,l})t^{R,kl}_{i,j}\alpha^k_{1,i}\alpha^l_{2,j} \\ &= \sup_{R\in F} \sup_{\substack{\alpha_1,\alpha_2\in\mathbb{R}^{nr} \\ \|\alpha_1\|_2\leq 1, \|\alpha_2\|_2\leq 1}} \frac{2}{n}\sum_{1\leq i,j\leq r}\sum_{1\leq k\leq n} R_{[i,j]}\alpha^k_{1,i}\alpha^k_{2,j} . \end{aligned}$$

From this last expression, it follows that E is a supremum of L_2 operator norms

$$E = \frac{2}{n} \sup_{R \in F} \varphi_{\max} \left(Diag^{(n)}(R) \right) ,$$

where $Diag^{(n)}(R)$ is the $(nr \times nr)$ block diagonal matrix such that each diagonal block is made of the matrix R. Since the largest eigenvalue of $Diag^{(n)}(R)$ is exactly the largest eigenvalue of R, we get

$$E = \frac{2}{n} \sup_{R \in F} \varphi_{\max}(R) . \tag{48}$$

Applying Proposition 8.3 and gathering identities (47) and (48) yields

$$\mathbb{P}(Z \geq \mathbb{E}(Z) + t) \leq \exp \left[- \left(\frac{t^2}{L_1 \mathbb{E}(V)} \bigwedge \frac{t}{L_2 B} \right) \right] ,$$

where $B = E$ and $V = D^2$. □

Proof of Lemma 8.2. This result is an extension of Corollary 4 in [BBLM05]. We shall closely follow the sketch of their proof adapting a few arguments. First, we upper bound the moments of $(T - \mathbb{E}(T))_+$. Then, we derive the deviation inequality from it. Here, $x_+ = \max(x, 0)$.

Lemma 8.4. *For all real numbers $q \geq 2$,*

$$\| (T - \mathbb{E}(T))_+ \|_q \leq \sqrt{Lq} \mathbb{E}(D) + LqE , \tag{49}$$

where $\|T\|_q^q$ stands for the q-th moment of the random variable T. The quantities D and E are defined in Lemma 8.2.

By Lemma 8.4, for any $t \geq 0$ and any $q \geq 2$,

$$\begin{aligned} \mathbb{P}\left(T \geq \mathbb{E}(T) + t\right) &\leq \frac{\mathbb{E}\left[(T - \mathbb{E}(T))_+^q\right]}{t^q} \\ &\leq \left(\frac{\sqrt{Lq}\mathbb{E}(D) + LqE}{t} \right)^q . \end{aligned}$$

The right-hand side is at most 2^{-q} if $\sqrt{Lq}\mathbb{E}(D) \leq t/4$ and $LqE \leq t/4$. Let us set

$$q_0 := \frac{t^2}{16L\mathbb{E}(D)^2} \wedge \frac{t}{4LE} .$$

If $q_0 \geq 2$, then $\mathbb{P}\left(T \geq \mathbb{E}(T) + t\right) \leq 2^{-q_0}$. On the other hand if $q_0 < 2$, then $4 \times 2^{-q_0} \geq 1$. It follows that

$$\mathbb{P}\left(T \geq \mathbb{E}(T) + t\right) \leq 4 \exp \left(- \frac{\log(2)}{4L} \left[\frac{t^2}{4\mathbb{E}(D)^2} \wedge \frac{t}{E} \right] \right) .$$

□

Proof of Lemma 8.4. This result is based on the entropy method developed in [BBLM05]. Let $f : \mathbb{R}^N \to \mathbb{R}$ be a measurable function such that $T = f(U_1, \dots, U_N)$. In the sequel, $U_1', \dots, U_N'$ denote independent copies of $U_1, \dots, U_N$. The random variable T_i' and V^+ are defined by

$$\begin{aligned} T_i' &:= f(U_1, \dots, U_{i-1}, U_i', U_{i+1}, \dots, U_N) \ , \\ V^+ &:= \mathbb{E}\left[\sum_{i=1}^N (T - T_i')_+^2 | U_1^N\right] \ , \end{aligned}$$

where U_1^N refers to the set $\{U_1, \dots, U_N\}$. Theorem 2 in [BBLM05] states that for any real $q \geq 2$,

$$\|(T - \mathbb{E}(T))_+\|_q \leq \sqrt{Lq}\|\sqrt{V^+}\|_q \ . \tag{50}$$

To conclude, we only have bound the moments of $\sqrt{V^+}$. By definition,

$$T = \sup_{t \in \mathcal{T}} \left| \sum_{\{i,j\}} U_i U_j t_{\{i,j\}} + \sum_{i=1}^N t_{\{i\}} + t_\varnothing \right| \ .$$

Since the set $\mathcal{T}$ is compact, this supremum is achieved almost surely at an element t^0 of $\mathcal{T}$. For any $1 \leq i \leq N$,

$$\begin{aligned} (T - T_i')_+^2 &\leq \left(\left| \sum_{\{k,l\}} U_k U_l t^0_{\{k,l\}} + \sum_{k=1}^N t^0_{\{k\}} + t^0_\varnothing \right| - \left| \sum_{\{k,l\}, k\neq i, l\neq i} U_i U_j t^0_{\{k,l\}} + \sum_{k\neq i} U_i' U_k t^0_{\{k,i\}} + \sum_{k=1}^N t^0_{\{k\}} + t^0_\varnothing \right| \right)_+^2 \\ &\leq \left((U_i - U_i') \left| \sum_{j \neq i} U_j t^0\{i,j\} \right| \right)^2 \ . \end{aligned}$$

Gathering this bound for any i between 1 and N, we get

$$\begin{aligned} V^+ &\leq \sum_{i=1}^N \mathbb{E}\left[\left((U_i - U_i') \left| \sum_{j\neq i} U_j t^0\{i,j\} \right| \right)^2 \middle| U_1^N \right] \\ &\leq 2\sum_{i=1}^N \left[\sum_{j\neq i} U_j t^0\{i,j\} \right]^2 \\ &\leq 2 \sup_{\alpha \in \mathbb{R}^N, \ \|\alpha\|_2 \leq 1} \left[\sum_{i=1}^N \alpha_i \left(\sum_{j\neq i} t^0_{\{i,j\}} U_j \right) \right]^2 \\ &\leq 2 \sup_{t\in\mathcal{T}} \sup_{\alpha \in \mathbb{R}^N, \ \|\alpha\|_2 \leq 1} \sum_{i=1}^N \left[U_i \sum_{j\neq i} \alpha_j t_{\{i,j\}} \right]^2 = 2D^2 \ . \end{aligned}$$

Combining this last bound with (50) yields

$$\begin{aligned} \|(T - \mathbb{E}(T))_+\|_q &\leq \sqrt{Lq}\sqrt{2}\|D\|_q \\ &\leq \sqrt{Lq}\left[\mathbb{E}(D) + \|(D - \mathbb{E}(D))_+\|_q\right] \ . \end{aligned} \tag{51}$$

Since the random variable D defined in Lemma 8.2 is a measurable function f_2 of the variables $U_1, \dots, U_N$, we apply again Theorem 2 in [BBLM05].

$$\|(D - \mathbb{E}(D))_+\|_q \leq \sqrt{Lq}\|\sqrt{V_2^+}\|_q,$$

where V_2^+ is defined by

$$V_2^+ := \mathbb{E}\left[\sum_{i=1}^N (D - D_i')_+^2 \middle| U_i^N\right],$$

and $D_i' := f_2(U_1, \dots, U_{i-1}, U_i', U_{i+1}, \dots, U_N)$. As previously, the supremum in D is achieved at some random parameter (t^0, α^0). We therefore upper bound V_2^+ as previously.

$$\begin{aligned} V_2^+ &\leq \sum_{i=1}^N \mathbb{E}\left[\left((U_i - U_i')\left(\sum_{j\neq i}\alpha_j^0 t_{\{i,j\}}^0\right)\right)^2 \middle| U_1^N\right] \\ &\leq 2\sum_{i=1}^N \left(\sum_{j\neq i}\alpha_j^0 t_{\{i,j\}}^0\right)^2 \\ &\leq 2\sup_{\alpha^{(2)}\in\mathbb{R}^N, \|\alpha\|_2\leq 1}\left(\sum_{i=1}^N \alpha_j^{(2)}\sum_{j\neq i}\alpha_i^0 t_{\{i,j\}}\right)^2 = 2E^2 . \end{aligned}$$

Gathering this upper bound with (51) yields

$$\|(T - \mathbb{E}(T))_+\|_q \leq \sqrt{Lq}\mathbb{E}(D) + LqE .$$

□

Proof of Lemma 8.3. We shall apply the central limit theorem in order to transfer results for Rademacher chaos to Gaussian chaos. Let f be the unique function satisfying $T = f(y_1, \dots, y_N)$ for any $(y_1, \dots, y_N) \in \mathbb{R}^N$. As the set $\mathcal{T}$ is compact, the function f is known to be continuous. Let $(U_i^{(j)})_{1\leq i\leq N, j\geq 0}$ an i.i.d. family of Rademacher variables. For any integer $n > 0$, the random variables $Y^{(n)}$ and $T^{(n)}$ are defined by

$$\begin{aligned} Y^{(n)} &:= \left(\sum_{j=1}^n \frac{U_1^{(j)}}{\sqrt{n}}, \dots, \sum_{j=1}^n \frac{U_N^{(j)}}{\sqrt{n}}\right) , \\ T^{(n)} &:= f\left(Y^{(n)}\right) . \end{aligned}$$

Clearly, $T^{(n)}$ is a supremum of Rademacher chaos of order 2 with nN variables and a constant term. By the central limit theorem, $T^{(n)}$ converges in distribution towards T as n tends to infinity. Consequently, deviation inequalities for the variables $T^{(n)}$ transfer to T as long as the quantities $\mathbb{E}\left[D^{(n)}\right]$, $E^{(n)}$, and $\mathbb{E}(T^{(n)})$ converge.

We first prove that the sequence $T^{(n)}$ converges in expectation towards T. As $T^{(n)}$ converges in distribution, it is sufficient to show that the sequence $T^{(n)}$ is asymptotically uniformly

integrable. The set $\mathcal{T}$ is compact, thus there exists a positive number t_∞ such that

$$\begin{aligned} T^{(n)} &\leq t_\infty \left[\sum_{i,j} |Y_i^{(n)} Y_j^{(n)}| + 1 \right] \\ &\leq t_\infty \left[1 + (N+1)/2 \sum_{i=1}^{N} \left(Y_i^{(n)} \right)^2 \right] . \end{aligned}$$

It follows that

$$\left(T^{(n)} \right)^2 \leq t_\infty^2 \left(\frac{N+1}{2} \right)^2 \frac{N+2}{2} \left[1 + \sum_{i=1}^{N} \left(Y_i^{(n)} \right)^4 \right] . \tag{52}$$

The sequence $Y_i^{(n)}$ does not only converge in distribution to a standard normal distribution but also in moments (see for instance [Bil95] p.391). It follows that $\overline{\lim}\mathbb{E}\left[\left(T^{(n)} \right)^2 \right] \leq \infty$ and the sequence $f\left(Y^{(n)} \right)$ is asymptotically uniformly integrable. As a consequence,

$$\lim_{n\to\infty} \mathbb{E}\left[T^{(n)} \right] = \mathbb{E}[T] .$$

Let us turn to the limit of $\mathbb{E}\left[D^{(n)} \right]$. As the variable $T^{(n)}$ equals

$$T^{(n)} = \sup_{t\in\mathcal{T}} \left| \sum_{\{i,j\}} t_{\{i,j\}} \sum_{1\leq k,l\leq n} \frac{U_i^{(k)} U_j^{(l)}}{n} + \sum_i t_i \sum_{1\leq k\leq n} \frac{U_i^{(k)}}{\sqrt{n}} \sum_{l\neq k} \frac{U_i^{(l)}}{\sqrt{n}} + t_\varnothing + \sum_i t_i \right| ,$$

it follows that

$$\begin{aligned} D^{(n)} &= \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nN},\ \|\alpha\|_2\leq 1} \left| \sum_{1\leq i\leq N} \sum_{1\leq k\leq n} U_i^{(k)} \left\{ \sum_{j\neq i} \frac{t_{\{i,j\}}}{n} \sum_{1\leq l\leq n} \alpha_j^{(l)} + 2 \sum_{l\neq k} 2 \frac{t_{\{i\}}}{n} \alpha_i^{(l)} \right\} \right| \\ &\leq \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nN},\ \|\alpha\|_2\leq 1} \left\{ \sum_i \frac{U_i^{(k)}}{\sqrt{n}} \sum_j (1+\delta_{i,j}) t_{\{i,j\}} \frac{\sum_{1\leq l\leq n} \alpha_j^{(l)}}{\sqrt{n}} \right\} + A^{(n)} , \end{aligned} \tag{53}$$

where the random variable $A^{(n)}$ is defined by

$$A^{(n)} := \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nN},\ \|\alpha\|_2\leq 1} \sum_{i=1}^{N} \sum_{j=1}^{n} t_{\{i\}} \frac{U_i^{(j)}}{n} \alpha_i^j .$$

Straightforwardly, one upper bounds $A^{(n)}$ by $\frac{t_\infty}{n} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{n} \left(U_i^{(j)} \right)^2}$ and its expectation satisfies

$$\mathbb{E}\left(\left| A^{(n)} \right| \right) \leq t_\infty \sqrt{\frac{N}{n}} ,$$

which goes to 0 when n goes to infinity. Thus, we only have to upper bound the expectation of the first term in (53). Clearly, the supremum is achieved only when for all $1 \leq j \leq N$, the sequence

$(\alpha_j^{(l)})_{1\leq l\leq n}$ is constant. In such a case, the sequence $(\alpha_j^{(1)})_{1\leq j\leq N}$ satisfies $\|\alpha^{(1)}\|_2 \leq 1/\sqrt{n}$. it follows that

$$\mathbb{E}\left[D^{(n)}\right] = \mathbb{E}\left\{\sup_{t\in\mathcal{T}}\sup_{\alpha\in\mathbb{R}^N\|\alpha\|_2\leq 1}\mathbb{E}\left[\sum_i Y_i^{(n)}\sum_j(1+\delta_{i,j})\alpha_j\right]\right\} + \mathcal{O}(\frac{1}{\sqrt{n}}) \ .$$

Let g be the function defined by $g\left(y_1,\ldots,y_N\right) = \sup_{t\in\mathcal{T}}\sup_{\alpha\in\mathbb{R}^N\|\alpha\|_2\leq 1}\left[\sum_i y_i\sum_j(1+\delta_{i,j})\alpha_j\right]$, for any $(y_1,\ldots,y_N)\in\mathbb{R}^N$. The function $g(.)$ is measurable and continuous as the supremum is taken over a compact set. As a consequence, $g(Y^{(n)})$ converges in distribution towards $g(Y)$. As previously, the sequence is asymptotically uniformly integrable since its moment of order 2 is uniformly upper bounded. It follows that $\lim\mathbb{E}\left[D^{(n)}\right] = \mathbb{E}\left[D\right]$.

Third, we compute the limit of $E^{(n)}$. By definition,

$$\begin{aligned} E^{(n)} &= \sup_{t\in\mathcal{T}}\sup_{\alpha_1,\alpha_2\in\mathbb{R}^{nN},\ \|\alpha_1\|_2\leq 1,\|\alpha_2\|_2\leq 1}\sum_{i=1}^N\sum_{k=1}^n\alpha_{1,i}^k\left[\sum_{j\neq i}\sum_{l=1}^n\alpha_{2,j}^{(l)}\frac{t_{\{i,j\}}}{n} + 2\sum_{l\neq k}\alpha_{2,i}^{(l)}\frac{t_{\{i\}}}{n}\right] \\ &= \sup_{t\in\mathcal{T}}\sup_{\alpha_1,\alpha_2,\ \|\alpha_1\|_2\leq 1,\|\alpha_2\|_2\leq 1}\sum_{i=1}^N\sum_{j=1}^N(1+\delta_{i,j})\frac{t_{\{i,j\}}}{n}\left[\sum_{k=1}^n\sum_{l=1}^n\alpha_{1,i}^{(k)}\alpha_{2,j}^{(l)}\right] + \mathcal{O}(\frac{1}{n}) \ . \end{aligned}$$

As for the computation of $D^{(n)}$, the supremum is achieved when the sequences $(\alpha_{1,i}^k)_{1\leq k\leq n}$ and $(\alpha_{2,j}^l)_{1\leq l\leq n}$ are constant for any $i\in\{1,\ldots,N\}$. Thus, we only have to consider the supremum over the vectors α_1 and α_2 in $\mathbb{R}^N$.

$$E^{(n)} = \sup_{t\in\mathcal{T}}\sup_{\alpha_1,\alpha_2\in\mathbb{R}^N\|\alpha_i\|_2\leq 1}\sum_{i=1}^N\sum_{j=1}^N(1+\delta_{ij})t_{i,j}\alpha_{1,i}\alpha_{2,j} + \mathcal{O}\left(\frac{1}{n}\right).$$

It follows that $E^{(n)}$ converges towards E when n tends to infinity.

The random variable $T^{(n)} - \mathbb{E}(T^{(n)})$ converges in distribution towards $T - \mathbb{E}(T)$. By Lemma 8.2 ,

$$\mathbb{P}(T - \mathbb{E}(T) \geq x) \leq \underline{\lim}\ \exp\left(-\frac{x^2}{\mathbb{E}[D^{(n)}]^2 L_1} \wedge \frac{x}{E^{(n)}L_2}\right) \ ,$$

for any $x > 0$. Combining this upper bound with the convergence of the sequences $D^{(n)}$ and $E^{(n)}$ allows to conclude. □

8.2 Proof of Theorem 3.1

Proof of Theorem 3.1. 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

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

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

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

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

$$l(\widetilde{\theta}_{\rho_1}, \theta) \leq l(\theta_{m,\rho_1}, \theta) + \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m}) \ . \tag{54}$$

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

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

satisfies

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

A similar lemma holds in the isotropic case. In particular, we choose $\alpha = \frac{K - K_0}{K_0}$ and $\delta = \sqrt{\frac{1+\alpha}{1+\alpha/2}}$. Lemma 8.5 implies that on the event Ω_ξ,

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

Thus, gathering this bound with inequality (54) yields

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

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

$$\begin{aligned} \frac{\delta(\alpha)^{1/2} - 1}{\delta(\alpha)^{1/2}} \mathbb{E}_\theta \left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right) \right] \leq \left[1 + \delta(\alpha)^{-1/2} \left(\delta(\alpha)^{1/2} - 1 \right)^{-1} \right] l\left(\theta_{m,\rho_1}, \theta\right) + \\ 2\mathrm{pen}(m) + \frac{\delta(\alpha)^2 L(\alpha)}{(\delta(\alpha) - 1)\left[\frac{\alpha^2}{1+\alpha/2} \wedge n \right]} \frac{\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} \ . \end{aligned} \tag{55}$$

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

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

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

Proof of Lemma 8.5. Throughout this proof, it is more convenient to express the quantities $\overline{\gamma}_{n,p}(.)$ and $l(.)$ in terms of covariance and precision matrices. Thanks to Equation (19), we also provide a matricial expression for $\gamma(.)$:

$$\gamma(\theta') = \frac{1}{p^2} tr\left[(I - C(\theta'))\,\Sigma\,(I - C(\theta'))\right] \ . \tag{56}$$

Gathering identities (56) and (17), we get

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \quad = \quad \frac{1}{p^2} tr\left[\left(\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2 - \left[I_{p^2} - C(\widetilde{\theta}_{\rho_1})\right]^2\right)\left(\overline{\mathbf{X^v X^{v*}}} - \Sigma\right)\right] \ .$$

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 8.15, they are jointly diagonalizable in the same orthogonal basis. In the sequel, P stands for an orthogonal matrix associated to this basis. Then, the matrices $C(\theta_{m,\rho_1})$, $C(\widetilde{\theta}_{\rho_1})$, and Σ respectively decompose in

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

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 $\mathbf{Y} := \sqrt{\Sigma^{-1}}\mathbf{X^v}$. Clearly, the components of $\mathbf{Y}$ follow independent standard normal distributions. Gathering these new notations, we get

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) = \frac{1}{p^2} tr\left[\left(\left[I_{p^2} - D(\theta_{m,\rho_1})\right]^2 - \left[I_{p^2} - D(\widetilde{\theta}_{\rho_1})\right]^2\right) D_\Sigma \left(\overline{\mathbf{YY}^*} - I_{p^2}\right)\right] \ . \tag{57}$$

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

Let $< .,. >_{\mathcal{H}}$ and $< .,. >_{\mathcal{H}'}$ be two inner products in the space of square matrices of size p^2 respectively defined by

$$< A, B >_{\mathcal{H}} := \frac{tr(A^* \Sigma B)}{p^2} \text{ and } < A, B >_{\mathcal{H}'} := \frac{tr(A^* D_\Sigma B)}{p^2} \ .$$

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

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

Besides, these two inner products clearly satisfy $\|C(\theta')\|_{\mathcal{H}} = \|D(\theta')\|_{\mathcal{H}'}$ for any $\theta' \in \Theta^+$. Gathering these new notations, we may upper bound (57) by

$$\begin{aligned} \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) \quad - \quad \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \leq \|[I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\|_{\mathcal{H}'} \times \\ \sup_{\substack{\theta_1 \in \Theta_m,\, \theta_2 \in \Theta_{\widehat{m}}, \\ \|[I_{p^2} - D(\theta_1)]^2 - [I_{p^2} - D(\theta_2)]^2\|_{\mathcal{H}'} \leq 1}} \left\langle \left[I_{p^2} - D(\theta_1)\right]^2 - \left[I_{p^2} - D(\theta_2)\right]^2, \left[\overline{\mathbf{YY}^*} - I_{p^2}\right] \right\rangle_{\mathcal{H}'} \end{aligned} \tag{58}$$

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

$$\begin{aligned} \|\left[I_{p^2} - D(\theta_{m,\rho_1})\right]^2 - \left[I_{p^2} - D(\widetilde{\theta}_{\rho_1})\right]^2\|_{\mathcal{H}'} &= tr\left[\left([I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\right)^2 \frac{D_\Sigma}{p^2}\right]^{\frac{1}{2}} \\ &= tr\left[\left[D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right]^2 \frac{D_\Sigma}{p^2}\left[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right]^2\right]^{1/2} \\ &\leq \varphi_{\max}\left[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right] \|D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\|_{\mathcal{H}'} \ . \end{aligned} \tag{59}$$

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

$$\varphi_{\max}\left[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right] = \varphi_{\max}\left[I_{p^2} - C(\theta_{m,\rho_1})\right] + \varphi_{\max}\left[I_{p^2} - C(\widetilde{\theta}_{\rho_1})\right] \leq 2\rho_1 \ .$$

Let us turn to the second term in (58). 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 $< \mid >_{\mathcal{H}'}$. Gathering these notations, we get

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

Applying the classical inequality $ab \leq \delta a^2 + \delta^{-1}b^2/4$ and gathering inequalities (58) and (59) yields

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \leq \delta^{-1}\|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|^2_{\mathcal{H}} + \rho_1^2\delta \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,\widehat{m}^2}} \frac{1}{p^2} tr^2\left[RD_\Sigma\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right] \ . \quad (60)$$

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

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

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

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

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 end of the section. Let us fix the positive numbers α and ξ. Applying Lemma 8.6 to any model $m' \in \mathcal{M}$, the event Ω'_ξ defined by

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

satisfies

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

From inequality (60), it follows that

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

conditionally to Ω'_ξ. By 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

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

By 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

$$\begin{aligned}\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \quad &\leq \quad \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) \\ &+ \frac{2\delta^2\rho_1^2\varphi_{\max}(\Sigma)}{np^2}\left[d_{m^2}(1+\alpha/2) + d_{\widehat{m}^2}(1+\alpha/2)\right] + \frac{8\xi^2\varphi_{\max}(\Sigma)\delta^2}{np^2(\delta-1)} \quad (61)\end{aligned}$$

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

Lemma 8.7. *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 end of this section. Defining the universal constant $K_0 := 2L$, we derive from (61) that

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

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

Proof of Lemma 8.6. We only consider here the anisotropic case, since the isotropic case is analogous. This result is based on the deviation inequality for suprema of Gaussian chaos of order 2 stated in Proposition 8.1. For any model m' belonging to $\mathcal{M}$, we shall upper bound the quantities $\mathbb{E}(Z_{m'})$, $B_{m'}$, and $\mathbb{E}(W_{m'})$ defined in (43).

1. Let us first consider the expectation of $Z_{m'}$. Let $U'_{m,m'}$ be the new vector space defined by

$$U'_{m,m'} := U_{m,m'} \frac{\sqrt{D_\Sigma}}{p} \ ,$$

where $U_{m,m'}$ is introduced in the proof of Lemma 8.5. This new space allows to handle the computation with the canonical inner product in the space of matrices. Let $\mathcal{B}^{(2)}_{m^2,m'^2}$ be the unit ball of $U'_{m,m'}$ with respect to the canonical inner product. If R belongs to $U_{m,m'}$, then $\|R\|_{\mathcal{H}'} = \|R\frac{\sqrt{D_\Sigma}}{p}\|_F$, where $\|.\|_F$ stands for the Frobenius norm.

$$\begin{aligned} Z_{m'} &= \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}} \frac{1}{p^2} tr\left[RD_\Sigma(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2})\right] \\ &= \sup_{R \in \mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\frac{\sqrt{D_\Sigma}}{p}\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right] \qquad (62) \\ &= \|\Pi_{U'_{m,m'}} \frac{\sqrt{D_\Sigma}}{p}\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\|_F \ , \end{aligned}$$

where $\Pi_{U'_{m,m'}}$ refers to the orthogonal projection with respect to the canonical inner product onto the space $U'_{m,m'}$. Let $F_1, \ldots, F_{d_{m^2,m'^2}}$ denote an orthonormal basis of $U'_{m,m'}$.

$$\begin{aligned} \mathbb{E}(Z^2_{m'}) &= \sum_{i=1}^{d_{m^2,m'^2}} \mathbb{E}\left[tr^2\left(F_i\sqrt{\frac{D_\Sigma}{p^2}}\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right)\right] \\ &= \sum_{i=1}^{d_{m^2,m'^2}} \mathbb{E}\left[\sum_{j=1}^{p^2} F_{i[j,j]}\frac{\sqrt{D_{\Sigma[j,j]}}}{p}(\overline{\mathbf{Y}\mathbf{Y}^*}_{[j,j]} - 1)\right]^2 \\ &= \sum_{i=1}^{d_{m^2,m'^2}} \frac{2}{np^2} tr(F_i D_\Sigma F_i) \\ &\leq \sum_{i=1}^{d_{m^2,m'^2}} \frac{2\varphi_{\max}(D_\Sigma)}{np^2} = \frac{2d_{m^2,m'^2}\varphi_{\max}(\Sigma)}{np^2} \ . \end{aligned}$$

Applying Cauchy-Schwarz inequality, it follows that

$$\mathbb{E}(Z_{m'}) \leq \sqrt{\frac{2d_{m^2,m'^2}\varphi_{\max}(\Sigma)}{np^2}} \ . \qquad (63)$$

2. Using the identity (62), the quantity $B_{m'}$ equals

$$B_{m'} = \frac{2}{n} \sup_{R \in \mathcal{B}^{(2)}_{m^2,m'^2}} \varphi_{\max}\left(R\frac{\sqrt{D_\Sigma}}{p}\right) \ .$$

As the operator norm is under-multiplicative and as it dominates the Frobenius norm, we get the following bound

$$B_{m'} \leq \frac{2\sqrt{\varphi_{\max}(\Sigma)}}{np} \ . \qquad (64)$$

3. Let us turn to bounding the quantity $\mathbb{E}(W_{m'})$. Again, by introducing the ball $\mathcal{B}^{(2)}_{m^2,m'^2}$, we get

$$\begin{aligned} W_{m'} &= \frac{4}{n} \sup_{R\in\mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}} \frac{1}{p^2} tr\left[R\overline{\mathbf{Y}\mathbf{Y}^*}D_\Sigma R\right] \\ &\leq \frac{4\varphi_{\max}(\Sigma)}{np^2} \sup_{R\in\mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\overline{\mathbf{Y}\mathbf{Y}^*}R\right] \\ &\leq \frac{4\varphi_{\max}(\Sigma)}{np^2}\left(1+\sup_{R\in\mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\left(\overline{\mathbf{Y}\mathbf{Y}^*}-I_{p^2}\right)R\right]\right) . \end{aligned}$$

Let $F_1,\dots F_{d_{m^2,m'^2}}$ an orthonormal basis of $U'_{m,m'}$ and let λ be a vector in $\mathbb{R}^{d_{m^2,m'^2}}$. We write $\|\lambda\|_2$ for its L_2 norm.

$$\begin{aligned} \mathbb{E}\left(\sup_{R\in\mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\left(\overline{\mathbf{Y}\mathbf{Y}^*}-I_{p^2}\right)R\right]^2\right) &= \mathbb{E}\left(\sup_{\|\lambda\|_2\leq 1}\sum_{i,j=1}^{d_{m^2,m'^2}}\lambda_i\lambda_j tr\left[F_iF_j(\overline{\mathbf{Y}\mathbf{Y}^*}/n-I_{p^2})\right]\right)^2 \\ &\leq \sum_{i,j=1}^{d_{m^2,m'^2}}\mathbb{E}\left(tr\left[F_iF_j(\mathbf{Y}\mathbf{Y}^*/n-I_{p^2})\right]^2\right) . \end{aligned}$$

The second inequality is a consequence of Cauchy-Schwarz inequality in $\mathbb{R}^{(d_{m^2,m'^2})^2}$ since the l_2 norm of the vector $(\lambda_i\lambda_j)_{1\leq i,j\leq d_{m^2,m'^2}}\in\mathbb{R}^{d^2_{m^2,m'^2}}$ is bounded by 1. Since the matrices F_i are diagonal, we get

$$\mathbb{E}\left(\sup_{R\in\mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R(\mathbf{Y}\mathbf{Y}^*/n-I)R\right]^2\right)\leq\frac{2}{n}\sum_{i,j=1}^{d_{m^2,m'^2}}\|F_iF_j\|_2^2 .$$

It remains to bound the norm of the products F_iF_j for any i,j between 1 and d_{m^2,m'^2}.

$$\sum_{i,j=1}^{d_{m^2,m'^2}}\|F_iF_j\|_2^2 = \sum_{i,j=1}^{d_{m^2,m'^2}}\sum_{k=1}^{p^2}F_{i[k,k]}{}^2F_{j[k,k]}{}^2=\sum_{k=1}^{p^2}\left(\sum_{i=1}^{d_{m^2,m'^2}}F_{i[k,k]}{}^2\right)^2 .$$

For any $k\in\{1,\dots,p^2\}$, $\sum_{i=1}^{d_{m^2,m'^2}}F_{i[k,k]}{}^2\leq 1$ since $\left(F_1,\dots,F_{d_{m^2,m'^2}}\right)$ form an orthonormal family. Hence, we get

$$\sum_{i,j=1}^{d_{m^2,m'^2}}\|F_iF_j\|_2^2 \leq \sum_{k=1}^{p^2}\sum_{i=1}^{d_{m^2,m'^2}}F_{i[k,k]}{}^2 = d_{m^2,m'^2} .$$

All in all, we have proved that

$$\mathbb{E}(W_{m'})\leq\frac{4\varphi_{\max}(\Sigma)}{np^2}\left[1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right] . \tag{65}$$

Gathering these three bounds and applying Proposition 8.1 allows to obtain the following deviation inequality:

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

As n and d_{m^2,m'^2} are larger than one, there exists a universal constant L_2' such that

$$\left[\frac{(\sqrt{1+\alpha/2}-1)^2 d_{m^2,m'^2}}{2L_1\left(1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right)} \bigwedge \frac{\sqrt{n}(\sqrt{1+\alpha/2}-1)\sqrt{d_{m^2,m'^2}}}{\sqrt{2}L_2}\right] \geq 4L_2'\sqrt{d_{m^2,m'^2}}\left[\left(\sqrt{1+\alpha/2}-1\right)^2 \wedge \left(\sqrt{1+\alpha/2}-1\right)\right].$$

Since the vector space $U_{m,m'}$ contains all the matrices $D(\theta')$ with θ' belonging to m', d_{m^2,m'^2} is larger than $d_{m'}$. Besides, by concavity of the square root function, it holds that $\sqrt{1+\alpha/2}-1 \geq \frac{\alpha}{4\sqrt{1+\alpha/2}}$. Setting $L_1' := \frac{1}{4L_1(1+\sqrt{2})} \wedge \frac{1}{\sqrt{2}L_2}$ and arguing as previously leads to

$$\left[\frac{\xi(\sqrt{1+\alpha/2}-1)\sqrt{d_{m^2,m'^2}}}{L_1\left(1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right)} \bigwedge \frac{\sqrt{n}\xi}{\sqrt{2}L_2}\right] \geq L_1'\xi\left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n}\right].$$

Gathering these two inequalities allows us to conclude that

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

□

Proof of Lemma 8.7. The approach falls in two parts. First, we relate the dimensions d_m and d_{m^2} to the number of nodes of the torus Λ that are closer than r_m or $2r_m$ to the origin $(0,0)$. We recall that the quantity r_m is introduced in Definition 2.1. Second, we compute a nonasymptotic upper bound of the number of points in $\mathbb{Z}^2$ that lie in the disc of radius r. This second step is quite tedious and will only give the main arguments.

Let m be a model of the collection $\mathcal{M}_1$. By definition, m is the set of points lying in the disc of radius r_m centered on $(0,0)$. Hence,

$$\Theta_m = vect\left\{\Psi_{i,j},\ (i,j) \in m\right\},$$

where the matrices $\Psi_{i,j}$ are defined by (14). As $\Psi_{i,j} = \Psi_{-i,-j}$, the dimension d_m of Θ_m is exactly the number of orbits of m under the action of the central symmetry s.

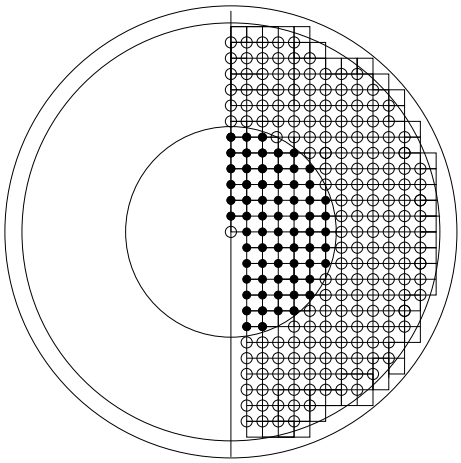

Figure 2: *The black dots represent the orbit space of m and the white dots represent the remaining points of the orbit space of $\mathcal{N}(m)$.*

As d_{m^2} is defined as the dimension of the space U_m, it also corresponds to the dimension of the space

$$vect\left\{C(\theta), \theta \in \Theta_m\right\} + vect\left\{C(\theta)^2, \theta \in \Theta_m\right\} \ , \tag{66}$$

which is clearly in one to one correspondence with U_m. Straightforward computations lead to the following identity:

$$C(\Psi_{i_1,j_1})C(\Psi_{i_2,j_2}) = C(\Psi_{i_1+i_2,j_1+j_2})\left[1 + s_{i_1+i_2,j_1+j_2}\right] + C(\Psi_{i_1-i_2,j_1-j_2})\left[1 + s_{i_1-i_2,j_1-j_2}\right]) \ ,$$

where $s_{x,y}$ is the indicator function of $x = -x$ and $y = -y$ in the torus Λ. Combining this property with the definition of Θ_m, we embed the space (66) in the space

$$vect\left\{C\left(\Psi_{i_1+i_2,j_1+j_2}\right),\ (i_1,j_1),(i_2,j_2) \in m \cup \{(0,0)\}\right\} \ ,$$

and this last space is in one to one correspondence with

$$vect\left\{\Psi_{i_1+i_2,j_1+j_2},\ (i_1,j_1),(i_2,j_2) \in m \cup \{(0,0)\}\right\} \ . \tag{67}$$

In the sequel, $\mathcal{N}(m)$ stands for the set $\{(i_1+i_2, j_1+j_2),\ (i_1,j_1),(i_2,j_2) \in m \cup \{(0,0)\}\}$. Thus, the dimension d_{m^2} is smaller or equal to the number of orbits of $\mathcal{N}(m)$ under the action of the symmetry s.

To conclude, we have to compare the number of orbits in m and the number of orbits in $\mathcal{N}(m)$. We distinguish two cases depending whether $2r_m + 1 \leq p$ or $2r_m + 1 > p$. First, we assume that $2r_m + 1 \leq p$. For such values the disc of radius r_m centered on the points $(0,0)$ in not overlapping itself on the torus except on a set of null Lebesgue measure. In the sequel, $\lfloor x \rfloor$ refers to the largest integer smaller than x. We represent the orbit space of m as in Figure 2. To any of these points, we associate a square of size 1. If we add $2 + 2\lfloor r_m \rfloor$ squares to the d_m first squares, we remark that the half disc centered on $(0,0)$ and with length r_m is contained in the reunion of these squares. Then, we get

$$d_m + 2 + 2\lfloor r_m \rfloor \geq \frac{\pi r_m^2}{2} \ . \tag{68}$$

The points in $\mathcal{N}(m)$ are closer than $2r_m$ from the origin. Consequently, all the squares associated to representants of $\mathcal{N}(m)$ are included in the disc of radius $2r_m + \sqrt{2}$.

$$d_{m^2} + 2 + 2\lfloor 2r_m \rfloor \leq \frac{\pi}{2}\left\{2r_m + \sqrt{2}\right\}^2 .$$

Combining these two inequalities, we are able to upper bound d_{m^2}

$$\begin{aligned} 2 + 2\lfloor 2r_m \rfloor + d_{m^2} &\leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 (d_m + 1 + 2\lfloor r_m \rfloor) , \\ d_{m^2} &\leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 d_m + 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 (1 + 2\lfloor r_m \rfloor) . \end{aligned}$$

Applying again inequality (68), we upper bound r_m:

$$r_m \leq \frac{2}{\pi}\left[1 + \sqrt{1 + \frac{\pi}{2}(1 + d_m)}\right] .$$

Gathering these two last bounds yields

$$d_{m^2} \leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 \left[1 + \frac{1}{d_m}\left(1 + \frac{4}{\pi}\left[1 + \sqrt{1 + \frac{\pi}{2}(1 + d_m)}\right]\right)\right] d_m .$$

This upper bound is equivalent to $4d_m$, when d_m goes to infinity. Computing the ratio d_{m^2}/d_m for every model m of small dimension allows to conclude.

Let us turn to the case $2r_m + 1 > p$. Suppose that p is larger or equal to 9. The lower bound (68) does not necessarily hold anymore. Indeed, the disc is overlapping with itself because of toroidal effects. Nevertheless, we obtain a similar lower bound by replacing r_m by $(p-1)/2$:

$$d_m + 2 + 2\lfloor \frac{p-1}{2} \rfloor \geq \frac{\pi(p-1)^2}{8}.$$

The number of orbits of Λ under the action of the symmetry s is $\frac{p^2+1}{2}$ if p is odd and $\frac{(p+1)^2-1}{2}$ if p is even. It follows that $d_{m^2} \leq \frac{(p+1)^2-1}{2}$. Gathering these two bounds, we get

$$\frac{d_{m^2}}{d_m} \leq \frac{(p+1)^2}{\pi(p-1)^2/4 - 2(p+1)} .$$

This last quantity is smaller than 4 for any $p \geq 9$. An exhaustive computation of the ratios when $p < 9$ allows to conclude.

Let us turn to the isotropic case. Arguing as previously, we observe that the dimension d_m^{iso} is the number of orbits of the set m under the action of the group G introduced in Definition 5 whereas d_{m^2} is smaller or equal to the number of orbits of $\mathcal{N}^{\text{iso}}(m)$ under the action of G. As for anisotropic models, we choose represent these orbits on the torus and associate squares of size 1 (see Figure 3). Assuming that $r_m < (p-1)/2$, we bound d_m and d_{m^2}.

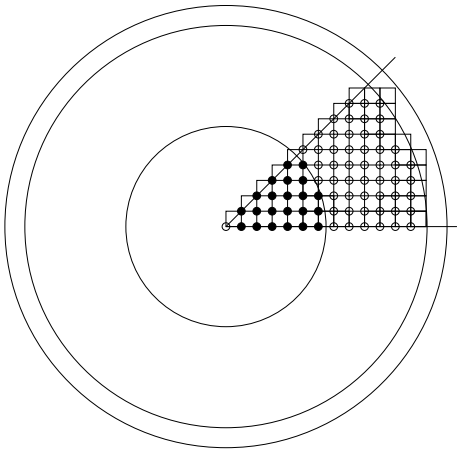

Figure 3: *The black dots represent the orbit space of m under the action of G and the white dots represent the remaining points of the orbit space of $\mathcal{N}^{iso}(m)$.*

$$\begin{aligned} d_m + 1 &\geq \frac{1}{8}\pi r_m^2 + \frac{1}{2}\lfloor \frac{\sqrt{2}r_m}{2} \rfloor \ , \\ d_{m^2} &\leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 \frac{1}{8}\pi r_m^2 + \frac{1}{2}\lfloor \sqrt{2}r_m \rfloor \ . \end{aligned}$$

Gathering these two inequalities, we get

$$d_{m^2} \leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 d_m \ .$$

As a consequence, d_{m^2} is smaller than $4d_m$ when d_m goes to infinity. As previously, computing the ratio d_{m^2}/d_m for models m of small dimension allows to conclude. The case $r_m > (p-1)/2$ is handled as for the anisotropic case. □

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

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

where $\mathrm{Co}\left[\mathcal{C}_m(\theta', r)\right]$ denotes the convex hull of $C_m(\theta', r)$. Similarly, let m be a model in $\mathcal{M}_1$ such $d_m^{\mathrm{iso}} \leq \sqrt{np}$ 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 $(1 - \|\theta'\|_1 - 8rd_m^{\mathrm{iso}})$ is positive,

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

Proof of Proposition 6.2. The first result derives from Lemma 8.8 applied to the hypercube $\mathcal{C}_m(0_p, \frac{1}{\sqrt{np^2}})$. We prove the second result using the same lemma with $\mathcal{C}_m(\theta', \frac{1-\|\theta\|_1}{\sqrt{n}p})$. □

Proof of Lemma 8.8. This lower bound is based on an application of Fano's approach. See [Yu97] 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.

Lemma 8.9. *The Kullback-Leibler entropy between two mean zero-Gaussian vectors of size p^2 with precision matrices $\left[I_{p^2} - C(\theta_1)\right]/\sigma^2$ and $\left[I_{p^2} - C(\theta_2)\right]/\sigma^2$ equals*

$$\mathcal{K}(\theta_1, \theta_2) = 1/2 \left[\log\left(\frac{|I_{p^2} - C(\theta_1)|}{|I_{p^2} - C(\theta_2)|} \right) + tr\left(\left[I_{p^2} - C(\theta_2)\right] \left[I_{p^2} - C(\theta_1)\right]^{-1} \right) - p^2 \right] ,$$

where for any square matrix A, $|A|$ refers to the determinant of A.

This statement is classical and its proof is omitted. The matrices $\left[I_{p^2} - C(\theta_1)\right]$ and $\left[I_{p^2} - C(\theta_2)\right]$ are diagonalizable in the same basis since they are symmetric block circulant (Lemma 8.15). Transforming vectors of size p^2 into $p \times p$ matrices, we respectively define λ_1 and λ_2 as the $p \times p$ matrices of eigenvalues of $\left[I_{p^2} - C(\theta_1)\right]$ and $\left[I_{p^2} - C(\theta_2)\right]$. It follows that

$$\mathcal{K}(\theta_1, \theta_2) = 1/2 \sum_{1 \le i,j \le p} \left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} - \log\left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} \right) - 1 \right) .$$

For any $x > 0$, the following inequality holds

$$x - 1 - \log(x) \le \frac{9}{64} \left(x - \frac{1}{x} \right)^2 .$$

It is easy to establish by studying the derivative of corresponding functions. As a consequence,

$$\begin{aligned} \frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} - \log\left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} \right) - 1 &\le \frac{9}{64} \left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} - \frac{\lambda_{1[i,j]}}{\lambda_{2[i,j]}} \right)^2 \\ &\le \frac{9}{64} \left(\frac{1}{\lambda_{1[i,j]}} + \frac{1}{\lambda_{2[i,j]}} \right)^2 \left(\lambda_{1[i,j]} - \lambda_{2[i,j]} \right)^2 . \end{aligned} \tag{69}$$

Let us first consider the anisotropic case. Let m be a model in $\mathcal{M}_1$ and let θ' belong $\Theta_m \cap \mathcal{B}_1(0_p, 1)$. We also consider a positive radius r such that $(1 - \|\theta'\|_1 - 2rd_m)$ is positive. For any θ_1, θ_2 in $\mathcal{C}_m(\theta', r)$ the matrices $(I_{p^2} - C(\theta_1))$ and $(I_{p^2} - C(\theta_2))$ are diagonally dominant and their eigenvalues $\lambda_{1[i,j]}$ and $\lambda_{2[i,j]}$ are larger than $1 - \|\theta'\|_1 - 2rd_m$.

$$\begin{aligned} \mathcal{K}(\theta_1, \theta_2) &\le \frac{9}{16(1 - \|\theta'\|_1 - 2rd_m)^2} \sum_{1 \le i,j \le p} \left(\lambda_{1[i,j]} - \lambda_{2[i,j]} \right)^2 \\ &\le \frac{9}{16(1 - \|\theta'\|_1 - 2rd_m)^2} \|C(\theta_1) - C(\theta_2)\|_F^2 \\ &\le \frac{9 d_m r^2 p^2}{8(1 - \|\theta'\|_1 - 2rd_m)^2} . \end{aligned} \tag{70}$$

We recall that $\|.\|_F$ refers to the Frobenius norm in the space of matrices.

Let us state Birgé's version of Fano's lemma [Bir05] and a combinatorial argument known under the name of Varshamov-Gilbert's lemma. These two lemma are taken from [Mas07] and respectively correspond to Corollary 2.18 and Lemma 4.7.

Lemma 8.10. (Birgé's lemma) *Let (S,d) be some pseudo-metric space and $\{\mathbb{P}_s, s \in S\}$ be some statistical model. Let κ denote some absolute constant smaller than one. Then for any estimator $\widehat{s}$ and any finite subset T of S, setting $\delta = min_{s,t\in T, s\neq t} d(s,t)$, provided that $\max_{s,t\in T} \mathcal{K}(\mathbb{P}_s, \mathbb{P}_t) \leq \kappa \log|T|$, the following lower bound holds for every $p \geq 1$,*

$$\sup_{s\in S} \mathbb{E}_s[d^p(s,\widehat{s})] \geq 2^{-p}\delta^p(1-\kappa) \ .$$

Lemma 8.11. (Varshamov-Gilbert's lemma) *Let $\{0,1\}^d$ be equipped with Hamming distance d_H. There exists some subset Φ of $\{0,1\}^d$ with the following properties*

$$d_H(\phi,\phi') > d/4 \text{ for every } (\phi,\phi') \in \Phi^2 \text{ with } \phi \neq \phi' \text{ and } \log|\Phi| \geq \frac{d}{8} \ .$$

Applying Lemma 8.10 with Hamming distance d_H and the set Φ introduced in Lemma 8.11 yields

$$\sup_{\theta\in\mathcal{C}_m(\theta',r)} \mathbb{E}_\theta\left[d_H\left(\widehat{\theta},\theta\right)\right] \geq \frac{d_m}{8}(1-\kappa) \ , \tag{71}$$

provided that

$$\frac{9d_m r^2 p^2 n}{8(1-\|\theta'\|_1 - 2rd_m)^2} \leq \frac{\kappa d_m}{8} \ . \tag{72}$$

Let us express (71) in terms of the Frobenius $\|.\|_F$ norm.

$$\sup_{\theta\in\mathcal{C}_m(\theta',r)} \mathbb{E}_\theta\left[\|C(\widehat{\theta}) - C(\theta)\|_F^2\right] \geq \frac{d_m r^2 p^2}{4}(1-\kappa) \ .$$

Since for every θ in the hypercube, $\sigma^{-2}(I_{p^2} - C(\theta))$ is diagonally dominant, its largest eigenvalue is smaller than $2\sigma^{-2}$. The loss function $l(\widehat{\theta},\theta)$ equals $\frac{\sigma^2}{p^2} tr\{[C(\widehat{\theta}) - C(\theta)](I - C(\theta))^{-1}[C(\widehat{\theta}) - C(\theta)]\}$. It follows that

$$\sup_{\theta\in\mathcal{C}_m(\theta',r)} \mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] \geq \sigma^2 \frac{d_m r^2}{8}(1-\kappa) \ . \tag{73}$$

Condition (72) is equivalent to $\frac{r^2}{(1-\|\theta'\|_1 - 2rd_m)^2} \leq \frac{\kappa}{9p^2 n}$. If we assume that

$$r^2 \leq \frac{\kappa\left(1-\|\theta'\|_1\right)^2}{18p^2 n} \ , \tag{74}$$

then $1 - \|\theta'\|_1 - 2rd_m \geq (1-\|\theta'\|_1)\left(1 - 2d_m\sqrt{\frac{\kappa}{18np^2}}\right)$. This last quantity is larger than $(1-\|\theta'\|_1)/\sqrt{2}$ if d_m is smaller than $\frac{3}{2}(\sqrt{2}-1)\sqrt{np^2/\kappa}$. Gathering inequality (73) and condition (74), we get

the lower bound

$$\begin{aligned}\inf_{\widehat{\theta}} \sup_{\theta \in \mathrm{Co}[\mathcal{C}_m(\theta',r)]} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] &\geq \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{C}_m\left[\theta', r\wedge(1-\|\theta'\|_1)\sqrt{\frac{\kappa}{18p^2n}}\right]} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] \\ &\geq L\left(r^2 \wedge \frac{\left(1-\|\theta'\|_1\right)^2}{np^2}\right) d_m\sigma^2 \ .\end{aligned}$$

One handles models of dimension d_m between $\frac{3}{2}(\sqrt{2}-1)\sqrt{np^2/\kappa}$ and $\sqrt{n}p$ by changing the constant L in the last lower bound.

Let us turn to sets of isotropic GMRFs. The proof is similar to the non-isotropic case, except for a few arguments. Let m belongs to the collection $\mathcal{M}_1$ and let θ' be an element of $\Theta_m^{\mathrm{iso}} \cap \mathcal{B}_1(0_p,1)$. Let r be such that $1-\|\theta'\|_1 - 8d_m^{iso}$ is positive. If θ_1 and θ_2 belong to the hypercube $\mathcal{C}_m^{\mathrm{iso}}(\theta',r)$, then

$$\mathcal{K}(\theta_1,\theta_2) \leq \frac{9d_m r^2 p^2}{2(1-\|\theta'\|_1 - 8rd_m^{\mathrm{iso}})^2} \ .$$

Applying Lemma 8.10 and 8.11, it follows that

$$\inf_{\widehat{\theta}} \sup_{\theta \in C_m^{\mathrm{iso}}(\theta',r)} \mathbb{E}_\theta \left[d_H\left(\widehat{\theta},\theta\right)\right] \geq \frac{d_m^{\mathrm{iso}}}{8}(1-\kappa) \ ,$$

provided that $\frac{9d_m r^2 p^2 n}{2(1-\|\theta'\|_1-8rd_m^{\mathrm{iso}})^2} \leq \frac{\kappa d_m^{\mathrm{iso}}}{8}$. As a consequence,

$$\inf_{\widehat{\theta}} \sup_{\theta \in C_m^{\mathrm{iso}}(\theta',r)} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] \geq \frac{d_m^{\mathrm{iso}} r^2}{8}(1-\kappa) \ ,$$

if $\frac{r^2}{(1-\|\theta'\|_1-8rd_m^{\mathrm{iso}})^2} \leq \frac{\kappa}{36p^2n}$. We conclude by arguing as in the isotropic case. □

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

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

If we choose

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

then $2rd_{m_i}$ is smaller than $1/8$ by assumption $(\mathbb{H}_a)$. Applying Lemma 8.12, we then derive that $\mathrm{Var}_\theta\left(X_{[0,0]}\right) \leq \sigma^2 + a_i^2$. Hence, we get the upper bound $\sum_{j=1}^{i}\left[\mathrm{Var}\left(X_{[0,0]}|X_{m_{j-1}}\right) - \mathrm{Var}\left(X_{[0,0]}|X_{m_j}\right)\right] \leq a_i^2$ and it follows that

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

since the sequence $(a_j)_{1\leq j\leq \mathrm{Card}(\mathcal{M}_1)}$ is non increasing. Consequently, $\mathrm{Co}\left[\mathcal{C}_m(0_p,r)\right]$ is a subset of $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1/2)$. By Lemma 8.8, we get

$$\begin{aligned}\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1/2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] &\geq L\sigma^2\left(\frac{a_i^2}{16\sigma^2}\wedge\frac{d_{m_i}}{np^2}\right)\\ &\geq L\left(a_i^2\wedge\frac{\sigma^2 d_{m_i}}{np^2}\right)\ . \end{aligned}\tag{75}$$

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

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1/2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] \geq L\sup_{i\leq\mathrm{Card}(\mathcal{M}_1),\ d_{m_i}\leq\sqrt{n}p}\left(a_i^2\wedge\frac{\sigma^2 d_{m_i}}{np^2}\right)\ .\tag{76}$$

If the maximal dimension $d_{m_{\mathrm{Card}(\mathcal{M}_1)}}$ is smaller than $\sqrt{n}p$, the proof is finished. In the opposite case, we need to show that the supremum (41) over all models $m\in\mathcal{M}_1$ is achieved at some model m of dimension less than $\sqrt{n}p$.

Lemma 8.13. *For any integer $1\leq i\leq Card(\mathcal{M}_1)-1$, the ratio $d_{m_{i+1}}/d_{m_i}$ is less than 2.*

Let i' be the largest integer such that $d_{m_{i'}}\leq\sqrt{n}p$. Since i' is smaller than $\mathrm{Card}(\mathcal{M}_1)$, we know from Lemma 8.13 that $\sqrt{n}p/2\leq d_{m_{i'}}\leq\sqrt{n}p$. By assumption $(\mathbb{H}_a)$, $a_{i'}^2$ is smaller than $\frac{\sigma^2}{d_{m_{i'}}}$. Gathering these bounds yields

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

Since the sequence $(a_i)_{1\leq i\leq\mathrm{Card}(\mathcal{M}_1)}$ is non increasing, the supremum (41) over all models in $\mathcal{M}_1$ is either achieved for some $i\leq i'$ or is smaller than $4\left(a_{i'}^2\wedge\frac{\sigma^2 d_{m_{i'}}}{np^2}\right)$. □

Proof of lemma 8.12. Let m be a model in $\mathcal{M}_1$, r be a positive number smaller than $\frac{1}{4d_m}$, and θ be an element of the convex hull of $\mathcal{C}_m(0_p,r)$. The covariance matrix of the vector X^v is $\Sigma=\sigma^2\left[I-C(\theta)\right]^{-1}$. Since the field X is stationary, $\mathrm{Var}_\theta(X_{[0,0]})$ equals any diagonal element of Σ. In particular, $\mathrm{Var}_\theta(X_{[0,0]})$ corresponds to the mean of the eigenvalues of Σ. The matrix $[I-C(\theta)]$ is block circulant. As in the proof of Lemma 71, we note λ the $p\times p$ matrix of the eigenvalues of $(I_{p^2}-C(\theta))$. By Lemma 8.15,

$$\lambda_{[i,j]}=1+\sum_{(k,l)\in\Lambda}\theta_{[k,l]}\cos\left[2\pi\left(\frac{ik}{p}+\frac{jl}{p}\right)\right]\ ,$$

for any $1\leq i,j\leq p$. Since θ belongs to the convex hull of $\mathcal{C}_m(0_p,r)$, $\theta_{[k,l]}$ is zero if $(k,l)\notin m$ and $|\theta_{[k,l]}|\leq r$ if $(k,l)\in m$. Thus $\sum_{(k,l)\in\Lambda}|\theta_{[k,l]}|$ is smaller than $1/2$. Applying Taylor-Lagrange inequality, we get

$$\frac{1}{1+x}\leq 1-x+\frac{x^2}{(1-|x|)^3}\ ,$$

for any x between -1 and 1. It follows that

$$\lambda_{[i,j]}^{-1}\leq 1-\sum_{k,l\in\Lambda}\theta_{[k,l]}\cos\left[2\pi\left(\frac{ik}{p}+\frac{jl}{p}\right)\right]+8\left\{\sum_{k,l\in\Lambda}\theta_{[k,l]}\cos\left[2\pi\left(\frac{ik}{p}+\frac{jl}{p}\right)\right]\right\}^2\ .\tag{77}$$

Summing this inequality for all $(i,j) \in \{1,\dots,p\}^2$, the first order term turns out to be $tr[C(\theta)]/p^2$ which is zero whereas the second term equals $8tr[C(\theta)^2]/p^2$. Since there are less than $2d_m$ non-zero terms on each line of the matrix $C(\theta)$, its Frobenius norm is smaller than $2d_m p^2 r^2$. Consequently, we obtain

$$\mathrm{Var}_\theta\left(X_{[0,0]}\right) \leq \sigma^2\left(1+16d_m r^2\right) \ .$$

□

Proof of Lemma 8.13. This property seems straightforward but the proof is a bit tedious. Let i be a positive integer smaller than $\mathrm{Card}(\mathcal{M}_1)$. By definition of the radius r_m in Equation (10), the model m_{i+1} is the set of nodes in $\Lambda \setminus \{(0,0)\}$ at a distance smaller or equal to $r_{m_{i+1}}$ from $(0,0)$, whereas the model m_i only contains the points in $\Lambda \setminus \{(0,0)\}$ at a distance strictly smaller than $r_{m_{i+1}}$ from the origin.

Let us first assume that $2r_{m_{i+1}} \leq p$. In such a case, the disc centered on $(0,0)$ with radius $r_{m_{i+1}}$ does not overlap with itself on the torus Λ. To any node in the neighborhood m_{i+1} and to the node $(0,0)$, we associate the square of size 1 centered on it. All these squares do not overlap and are included in the disc of radius $r_{m_{i+1}} + \sqrt{2}/2$. Hence, we get the upper bound $2d_{m_{i+1}} + 1 \leq \pi(r_{m_{i+1}} + \sqrt{2}/2)^2$. Similarly, the disc of radius $r_{m_{i+1}} - \sqrt{2}/2$ is included in the union of the squares associated to the nodes $m_i \cup \{0,0\}$. It follows that $2d_{m_i}+1$ is larger or equal to $\pi\left(r_{m_{i+1}} - \sqrt{2}/2\right)^2$. Gathering these two inequalities, we obtain

$$\frac{d_{m_{i+1}}}{d_{m_i}} \leq \frac{\left(r_{m_{i+1}} + \frac{\sqrt{2}}{2}\right)^2 - 1}{\left(r_{m_{i+1}} - \frac{\sqrt{2}}{2}\right)^2 - 1} \ ,$$

if $r_{m_{i+1}}$ is larger than $1+\sqrt{2}/2$. If $r_{m_{i+1}}$ larger than 5, this upper bound is smaller than two. An exhaustive computation for models of small dimension allows to conclude.

If $2r_{m_{i+1}} \geq p$ and $2r_{m_i} < p$, then the preceding lower bound of d_{m_i} and the preceding upper bound of $d_{m_{i+1}}$ still hold. Finally, let us assume that $2r_{m_i} \geq p$. Arguing as previously, we conclude that $2d_{m_i} + 1 \geq \pi(p/2 - \sqrt{2}/2)^2$. The largest dimension of a model $m \in \mathcal{M}_1$ is $(p^2-1)/2$ if p is odd and $((p+1)^2-3)/2$ if p is even. Thus, $d_{m_{i+1}} \leq \frac{(p+1)^2-3}{2}$. Gathering these two bounds yields

$$\frac{d_{m_{i+1}}}{d_{m_i}} \leq \frac{(p+1)^2-3}{\left(\frac{p}{2} - \frac{\sqrt{2}}{2}\right)^2} \ ,$$

which is smaller than 2 if p is larger than 10. Exhaustive computations for small p allow to conclude.

□

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 8.8, we get the following minimax lower bound

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

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

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

We conclude by combining the two different bounds. □

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.

Let $\mathcal{E}(a)$ be a pseudo-ellipsoid that satisfies Assumption $(\mathbb{H}_a)$ and such that $a_1^2\geq\frac{1}{np^2}$. For any θ in $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)$, the penalty term satisfies $\mathrm{pen}(m)=K\sigma^2\rho_1^2\rho_2 d_m/np^2$ is larger than $Kd_m\varphi_{\max}(\Sigma)/np^2$. Applying Theorem3.1, we upper bound the risk $\widetilde{\theta}_{\rho_1}$

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right]\leq L_1(K)\inf_{m\in\mathcal{M}_1}\left[l(\theta_{m,\rho_1},\theta)+\mathrm{pen}(m)\right]+L_2(K)\rho_2\frac{\sigma^2}{np^2}\ ,$$

for any $\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)$. It follows that

$$\sup_{\theta\in}\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right]\leq L(K)\inf_{m\in\mathcal{M}_1\ ,d_m>0}\left[l(\theta_{m,\rho_1},\theta)+\rho_1^2\rho_2\sigma^2\frac{d_m}{np^2}\right]\ .$$

Let i be a positive integer smaller or equal than $\mathrm{Card}(\mathcal{M}_1)$. We know from Section 4.1 that the bias $l(\theta_{m_i},\theta)$ of the model m_i equals $\mathrm{Var}(X_{[0,0]}|X_{m_i})-\sigma^2$. Since θ belongs to the set $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)$, the bias term is smaller or equal to a_{i+1}^2 with the convention $a^2_{\mathrm{Card}(\mathcal{M}_1)+1}=0$. Hence , the previous upper bound becomes

$$\begin{aligned}\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right] &\leq L(K)\inf_{1\leq i\leq\mathrm{Card}(\mathcal{M}_1)}\left[a_{i+1}^2+\rho_1^2\rho_2\sigma^2\frac{d_{m_i}}{np^2}\right]\\ &\leq L(K,\rho_1,\rho_2)\inf_{1\leq i\leq\mathrm{Card}(\mathcal{M}_1)}\left[a_{i+1}^2+\frac{\sigma^2 d_{m_i}}{np^2}\right]\ .\end{aligned}\tag{78}$$

Applying Proposition 6.6 to the set $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)$, we get

$$\begin{aligned}\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] &\geq \inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]\\ &\geq L\sup_{1\leq i\leq\mathrm{Card}(\mathcal{M}_1)}\left(a_i^2\wedge\sigma^2\frac{d_{m_i}}{np^2}\right)\ .\end{aligned}$$

Let us define i^* by

$$i^*:=\sup\left\{1\leq i\leq\mathrm{Card}(\mathcal{M}_1)\,,\,a_i^2\geq\frac{\sigma^2 d_{m_i}}{np^2}\right\}\ ,$$

with the convention $\sup\varnothing=0$. Since $a_1^2\geq\sigma^2/np^2$, i^* is larger or equal to one. It follows that

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,\eta)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]\geq L_2\left(a_{i^*+1}^2\vee\frac{\sigma^2 d_{m_{i^*}}}{np^2}\right)\ .$$

Meanwhile, the upper bound (78) on the risk of $\widetilde{\theta}_{\rho_1}$ becomes

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right]\leq L(K,\rho_1,\rho_2)\left(a_{i^*+1}^2+\frac{\sigma^2 d_{m_{i^*}}}{np^2}\right)\leq 2L(K,\rho_1,\rho_2)\left(a_{i^*+1}^2\vee\frac{\sigma^2 d_{m_{i^*}}}{np^2}\right)\ ,$$

which allows to conclude. □

8.4 Proofs of the asymptotic risk bounds

Proof of Proposition 4.4. This result is closely related to Proposition 4.11 in [Guy95]. In fact, we extend his proof to stationary fields on a torus. In the sequel, we shall only consider non-isotropic 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

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

For any $(i,j) \in \{1,\ldots,p\}^2$, 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 neighbours in the model m. Contrary to the previous proofs, we need to express the $n \times p^2$ matrix $\mathbf{X^V}$ in terms of a vector. This is why we define the vector $\mathbf{X^V}$ of size np^2 as

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

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

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

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

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. Afterwards, 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

$$\begin{aligned} \check{a}_m &:= \left(\left(\boldsymbol{\chi}_m^{\mathbf{V}}\right)^* \boldsymbol{\chi}_m^{\mathbf{V}} \right)^{-1} \boldsymbol{\chi}_m^{\mathbf{V}} \mathbf{X^V} \\ \check{\theta}_m &:= \sum_{k=1}^{d_m} \check{a}_m[k] \Psi_{i_k,j_k} \ , \end{aligned} \tag{79}$$

where we recall that $(\Psi_{i_1,j_1},\ldots,\Psi_{i_{d_m},j_{d_m}})$ is a basis of Θ_m. Obviously, $\check{\theta}_m$ is a Conditional least squares estimator since it minimizes the expression (16) of $\gamma_{n,p}(.)$ 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

$$X_{[k,l]} = \sum_{(i,j)\in m} \theta_{m,\rho_1[i,j]} X_{[k+i,l+j]} + \epsilon_m[k,l] \ , \tag{80}$$

where $\epsilon_m[k,l]$ 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

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

Gathering this last identity with (79) yields

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

where the vector $\boldsymbol{\epsilon}_m^{\mathbf{V}}$ of size np^2 corresponds to the n observations of the vector ϵ_m^v. When n goes to the infinity, $\frac{1}{np^2}\left(\boldsymbol{\chi}_m^{\mathbf{V}}\right)^* \boldsymbol{\chi}_m^{\mathbf{V}}$ converges almost surely to the covariance matrix V by the law of large numbers. By definition, the variable $\epsilon_{m[i,j]}$ is independent from the $(i-1)p+j$th row of $\chi_{m[i,j]}^v$. It follows that $\mathbb{E}_\theta\left(\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\epsilon}^{\mathbf{V}}\right)=0$. Applying again the law of large numbers we conclude that $\check{a}_m$ converges almost surely towards a_m and that $\check{\theta}_m$ converges almost surely towards θ_{m,ρ_1}. Besides, the central limit theorem states that the random vector $\frac{1}{\sqrt{np}}\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\epsilon}^{\mathbf{V}}$ converges in distribution towards a zero mean Gaussian vector whose covariance matrix equals $\frac{1}{p^2}\mathrm{Var}_\theta\left(\chi_m^v\epsilon_m^v\right)$. By decomposition (80), $\epsilon_m^v=(I-C(\theta_{m,\rho_1}))X^v$ while the k-th row of χ_m^v equals $[C(\Psi_{i_k,j_k})X^v]^*$. Thus, for any $1\le k,l\le d_m$,

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

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

$$\frac{1}{p^2}\mathrm{Var}_\theta\left(\chi_m^v\epsilon_m^v\right)_{[k,l]}=$$
$$\frac{2\sigma^4}{p^2}\mathrm{cov}_\theta\left[\left[I-C(\theta)\right]^{-1}C(\Psi_{i_k,j_k})\left[I-C(\theta_{m,\rho_1})\right]\left[I-C(\theta)\right]^{-1}C(\Psi_{i_l,j_l})\left[I-C(\theta_{m,\rho_1})\right]\right]\ . \quad (81)$$

By Lemma 8.15, all these matrices are diagonalizable in the same basis and therefore commute with each other. We conclude that $\frac{1}{p^2}\mathrm{Var}_\theta\left(\chi_m^v\epsilon_m^v\right)=2\sigma^4W$ and

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

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

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

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

$$l\left(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}\right)=\mathbb{E}_\theta\left[\sum_{k=1}^{d_m}\left(\widehat{a}_{m[k]}-a_{m[k]}\right)tr\left(\Psi_{i_k,j_k}X\right)\right]^2=tr\left[V\left(\widehat{a}_m-a_m\right)^*\left(\widehat{a}_m-a_m\right)\right]\ . \quad (82)$$

By Portmanteau's Lemma, $np^2l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})$ converges in distribution towards a random variable whose expectation is $tr\left(WV^{-1}\right)$. In order to conclude, it remains to prove that the sequence $\left[np^2l\left(\widehat{\theta}_{m,\rho_1},\theta\right)\right]_{n\ge1}$ 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 (55) the following property. For any $\xi>0$, with probability larger than $1-L_1\exp\left[-L_2\xi\right]$,

$$np^2l\left(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}\right)\le L_3d_m\varphi_{\max}(\Sigma)+L_4\xi^2\varphi_{\max}(\Sigma)\ .$$

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,

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

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

Proof of Corollary 4.6. For the sake of simplicity, we assume that for any node $(i,j) \in m$, the nodes (i,j) and $(-i,-j)$ are different in Λ. If this is not the case, we only have to slightly modify the proof in order to take account that $\|\Psi_{i,j}\|_F^2$ may equal one. The matrix V is the covariance of the vector of size d_m

$$\left(X_{i_1,j_1} + X_{-i_1,-j_1}, \ldots, X_{i_{d_m},j_{d_m}} + X_{-i_{d_m},-j_{d_m}}\right) \ . \tag{83}$$

Since the matrix Σ of X^v is positive, V is also positive. Moreover, its largest eigenvalue is larger than $2\varphi_{\max}(\Sigma)$.

Let us assume first the θ belongs to Θ_m^+ and that Assumption $(\mathbb{H}_1)$ is fulfilled. By the first result of Proposition 4.4,

$$\lim_{n\to+\infty} np^2\mathbb{E}\left[l\left(\widehat{\theta}_{m,\rho_1}, \theta\right)\right] = 2\sigma^4 tr\left[IL_m V^{-1}\right] \geq \frac{\sigma^4}{\varphi_{\max}(\Sigma)} tr[IL_m] = 2\sigma^4 \frac{d_m}{\varphi_{\max}(\Sigma)} \ ,$$

which corresponds to the first lower bound (30).

Let us turn to the second result. We now assume that θ satisfies Assumption $(\mathbb{H}_2)$. By the identity (28) of Proposition 4.4, we only have to lower bound the quantity $tr\left[VW^{-1}\right]$.

$$tr\left[V^{-1}W\right] \quad \geq \quad \varphi_{\max}(V)^{-1} tr\left[W\right] \geq \frac{1}{2\varphi_{\max}(\Sigma)} tr[W] \ .$$

Since the matrix $\Sigma^{-1} = \sigma^{-2}\left[I_{p^2} - C(\theta)\right]$ is diagonally dominant, its smallest eigenvalue is larger than $\sigma^{-2}(1 - \|\theta\|_1)$. The matrix $\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2 \left[I_{p^2} - C(\theta)\right]^{-2}$ is symmetric positive. It follows that W is also symmetric positive definite. Hence, we get

$$tr\left[V^{-1}W\right] \quad \geq \quad \frac{\sigma^{-2}}{2}\left[1 - \|\theta\|_1\right] \sum_{k=1}^{d_m} \frac{1}{p^2} tr\left[C(\Psi_{i_k,j_k})^2 \left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2 \left[I_{p^2} - C(\theta)\right]^{-2}\right] \tag{84}$$

The largest eigenvalue of $\left[I_{p^2} - C(\theta)\right]$ is smaller than 2 and the smallest eigenvalue of $\left[I_{p^2} - C(\theta_{m,\rho_1})\right]$ is larger than $1 - \|\theta_{m,\rho_1}\|_1$. By Lemma 8.15, these two matrices are jointly diagonalizable and the smallest eigenvalue of $\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2 \left[I_{p^2} - C(\theta)\right]^{-2}$ is therefore larger than $(1 - \|\theta_{m,\rho_1}\|_1)^2/4$. Gathering this lower bound with (84) yields

$$tr\left[V^{-1}W\right] \geq \frac{d_m \sigma^{-2}}{2}\left[1 - \|\theta\|_1\right]\left[1 - \|\theta_{m,\rho_1}\|_1\right]^2 \ .$$

Lemma 4.1 states that $\|\theta_{m,\rho_1}\|_1 \leq \|\theta\|_1$. Combining these two lower bounds enables to conclude. □

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 $\frac{\sigma^2}{1-\eta}$. Applying Theorem 3.1, we get

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

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

Proof of Example 4.8.

Lemma 8.14. *For any θ is the space $\Theta^{+,\mathrm{iso}}_{m_1}$, the asymptotic variance term of $\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1}$ equals*

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1}, \theta\right)\right] = 2\sigma^4\frac{tr\left(H^2\right)}{tr\left(H^2\Sigma\right)}\ .$$

If θ belongs to $\Theta^{+,\mathrm{iso}}$ and also satisfies $(\mathbb{H}_2)$, then

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}^{\mathrm{iso}}_{m_1,\rho_1}, \theta^{\mathrm{iso}}_{m_1,\rho_1}\right)\right] = 2\frac{tr\left\{\left[(I-\theta^{\mathrm{iso}}_{m_1,\rho_1}[1,0]H)H\Sigma\right]^2\right\}}{tr(H^2\Sigma)}\ , \tag{85}$$

where the $p^2\times p^2$ matrix H is defined as $H := C\left(\Psi^{\mathrm{iso}}_{1,0}\right)$.

Proof of Lemma 8.14. Apply Proposition 4.4 noting that $V = tr[H\Sigma H]/p^2$ and

$$W = \frac{tr\left\{\left[(I-\theta_{m_1^{\mathrm{iso}}}[1,0]H)H\Sigma\right]^2\right\}}{\sigma^4p^2}\ .$$

To prove the second result, we observe that $\Theta^{+,\mathrm{iso}}_{m_1}$ equals $\Theta^{+,\mathrm{iso}}_{m_1,2}$. It is stated for instance in Table 2. □

Since the matrix θ belongs to $\Theta^{+,\mathrm{iso}}_{m_1}$, we may apply the second result of Lemma 8.14. Straightforward computations lead to $tr(H^2) = \|C\left(\Psi^{\mathrm{iso}}_{1,0}\right)\|_F^2 = 4p^2$ and

$$tr(H^2\Sigma) = 4p^2\left[\mathrm{Var}(X_{[0,0]}) + 2\mathrm{cov}_\theta(X_{[0,0]}, X_{[1,1]}) + \mathrm{cov}_\theta\left(X_{[0,0]}, X_{[2,0]}\right)\right]\ .$$

Since the field X is an isotropic GMRF with four nearest neighbors,

$$X_{[0,0]} = \theta_{[1,0]}\left(X_{[1,0]} + X_{[-1,0]} + X_{[0,1]} + X_{[0,-1]}\right) + \epsilon_{[0,0]}\ ,$$

where $\epsilon_{[0,0]}$ is independent from every variable $X_{[i,j]}$ with $(i,j)\neq 0$. Multiplying this identity by $X_{[1,0]}$ and taking the expectation yields

$$\mathrm{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) = \theta_{[1,0]}\left[\mathrm{Var}\left(X_{[0,0]}\right) + 2\mathrm{cov}_\theta(X_{[0,0]}, X_{[1,1]}) + \mathrm{cov}_\theta\left(X_{[0,0]}, X_{[2,0]}\right)\right]\ .$$

Hence, we obtain $tr\left(H^2\Sigma\right) = 4\mathrm{cov}_\theta(X_{[0,0]}, X_{[1,0]})/\theta_{[1,0]}$ and

$$\frac{tr(H^2)}{tr(H^2\Sigma)} = \frac{\theta_{[1,0]}}{\mathrm{cov}_\theta(X_{[0,0]}, X_{[1,0]})}\ ,$$

which concludes the first part of the proof.

This second part is based on the spectral representation of the field X and follows arguments which come back to Moran [Mor73]. We shall compute the limit of $\mathrm{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right)$ when the size of Λ goes to infinity. As the field X is stationary on Λ, we may diagonalize its covariance matrix Σ applying Lemma 8.15. We note D_Σ the corresponding diagonal matrix defined by

$$D_{\Sigma[(i-1)p+j,(i-1)p+j]} = \sum_{k=1}^{p}\sum_{l=1}^{p} \mathrm{cov}_\theta\left(X_{[0,0]}, X_{[k,l]}\right) \cos\left[2\pi\left(\frac{ki}{p}+\frac{lj}{p}\right)\right] \ ,$$

for any $1 \leq i,j \leq p$. Straightforwardly, we express $\mathrm{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right)$ as a linear combination of the eigenvalues

$$\mathrm{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) = \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} \cos\left(2\pi\frac{i}{p}\right) D_{\Sigma[(i-1)p+j,(i-1)p+j]} \ .$$

Applying Lemma 8.15 to the matrix Σ^{-1} and noting that $\theta \in \Theta^{\mathrm{iso},+}$ allows to get another expression of the eigenvalues of Σ

$$D_{\Sigma[(i-1)p+j,(i-1)p+j]} = \frac{\sigma^2}{1-2\theta_{[1,0]}\left[\cos\left(\frac{2\pi i}{p}\right)+\cos\left(\frac{2\pi j}{p}\right)\right]} \ .$$

We then combine these expression. By symmetry between i and j we get

$$\mathrm{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) = \frac{\sigma^2}{2p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)}{1-2\theta_{[1,0]}\left[\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)\right]} \ .$$

If we let p go to infinity, this sum converges to the following integral

$$\begin{aligned}\lim_{p\to+\infty} \mathrm{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) &= \frac{\sigma^2}{2}\int_0^1\int_0^1 \frac{\cos(2\pi x)+\cos(2\pi y)}{1-2\theta_{[1,0]}\left(\cos(2\pi x)+\cos(2\pi y)\right)}dxdy \\ &= \frac{\sigma^2}{2\theta_{[1,0]}}\left[-1+\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\frac{1}{1-2\theta_{[1,0]}\left[\cos(x)+\cos(y)\right]}dxdy\right] \ .\end{aligned}$$

This last elliptic integral is asymptotically equivalent to $\frac{\log 16}{4(1-4\theta_{[1,0]})}$ when $\theta_{[1,0]} \to \frac{1}{4}$ as observed for instance by Moran [Mor73]. We conclude by substituting this limit in expression (33). □

Proof of Example 4.9. First, we compute $[\theta^{(p)}]^{\mathrm{iso}}_{m_1}{}_{[1,0]}$. By Lemma 4.1, it minimizes the function $\gamma(.)$ defined in (19) over the whole space $\Theta_{m_1^{\mathrm{iso}}}$. We therefore obtain

$$[\theta^{(p)}]^{\mathrm{iso}}_{m_1}{}_{[1,0]} = \frac{tr\left[\Sigma H\right]}{tr\left[\Sigma H^2\right]} \ .$$

Once again, we apply Lemma 8.15 to simultaneously diagonalize the matrices H and Σ^{-1}. As previously, we note D_Σ the corresponding diagonal matrix of Σ.

$$\begin{aligned}D_{\Sigma[(i-1)p+j,(i-1)p+j]} &= \frac{\sigma^2}{1-2\alpha\left[\cos\left(2\pi\left(\frac{pi}{4p}+\frac{pj}{4p}\right)\right)+\cos\left(2\pi\left(\frac{-pi}{4p}+\frac{pj}{4p}\right)\right)\right]} \\ &= \frac{\sigma^2}{1-4\alpha\cos\left(\pi\frac{i}{2}\right)\cos\left(\pi\frac{j}{2}\right)} \ .\end{aligned}$$

Analogously, we compute the diagonal matrix $D\left(\Psi_{1,0}^{\mathrm{iso}}\right)$

$$D\left(\Psi_{1,0}^{\mathrm{iso}}\right)_{[(i-1)p+j,(i-1)p+j]} = 2\left[\cos\left(2\pi\tfrac{i}{p}\right)+\cos\left(2\pi\tfrac{j}{p}\right)\right] .$$

Combining these two last expressions, we obtain

$$tr(H\Sigma) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^2\frac{2\left[\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)\right]}{1-4\alpha\cos\left(\pi\frac{i}{2}\right)\cos\left(\pi\frac{j}{2}\right)} .$$

Let us split this sum in 16 parts depending on the congruence of i and j modulo 4. As each if of these 16 sums is shown to be zero, we conclude that $tr(H\Sigma) = [\theta^{(p)}]^{\mathrm{iso}}_{m_1}{}_{[1,0]} = 0$. By Lemma 8.14, the asymptotic risk of $\widehat{\theta^{(p)}}^{\mathrm{iso},\rho_1}_{m_1}$ therefore equals

$$\lim_{n\to+\infty} np^2\mathbb{E}_{\theta^{(p)}}\left[l\left(\widehat{\theta^{(p)}}^{\mathrm{iso},\rho_1}_{m_1},[\theta^{(p)}]^{\mathrm{iso}}_{m_1}\right)\right] = \frac{tr(H^4\Sigma^2)}{tr(H^2\Sigma)} .$$

First, we lower bound the numerator

$$tr(H^4\Sigma^2) = \sigma^4\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\left\{2\left[\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)\right]\right\}^4}{\left\{1-4\alpha\cos\left(\pi\frac{i}{2}\right)\cos\left(\pi\frac{j}{2}\right)\right\}^2} .$$

As each term of this sum is non-negative, we may only consider the coefficients i and j which are congruent to 0 modulo 4.

$$tr(H^4\Sigma^2) \geq \sigma^4\sum_{i=0}^{p/4-1}\sum_{j=0}^{p/4-1}\frac{16\left[\cos\left(2\pi\frac{i}{p/4}\right)+\cos\left(2\pi\frac{j}{p/4}\right)\right]^4}{(1-4\alpha)^2} .$$

If we let go p to infinity, we get the lower bound

$$\lim_{p\to+\infty}\frac{tr(H^4\Sigma^2)}{p^2} \geq \frac{\sigma^4}{(1-4\alpha)^2}\int_0^1\int_0^1\left[\cos(2\pi x)+\cos(2\pi y)\right]^4 dxdy .$$

Similarly, we upper bound $tr(H^2\Sigma)$ and let p go to infinity

$$\lim_{p\to+\infty}\frac{tr(H^2\Sigma)}{p^2} \leq \frac{4\sigma^2}{1-4\alpha}\int_0^1\int_0^1\left[\cos(2\pi x)+\cos(2\pi y)\right]^2 dxdy .$$

Combining these two bounds allows to conclude

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

□

Appendix

Lemma 8.15. *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*

$$D(\theta)_{[(i-1)p+j,(i-1)p+j]} = \sum_{k=1}^{p}\sum_{l=1}^{p}\theta_{[k,l]}\cos\left(2\pi(ki/p+lj/p)\right) \tag{86}$$

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

It is proved as in [RH05] Sect.2.6.2 to the price of a slight modification in order to take into account the fact that P has is orthogonal and not unitary. 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 (86) explicitly provides the eigenvalues of the $C(\theta)$ as the two-dimensional discrete Fourier transform of the $p \times p$ matrix θ.

Proof of Lemma 1.1. Let θ be a $p \times p$ matrix that satisfies condition (3). For any $1 \leq i_1, i_2 \leq p$, we define the $p \times p$ submatrix C_{i_1,i_2} as

$$C_{i_1,i_2}[j_1,j_2] := C(\theta)_{[(i_1-1)p+j_1,(i_2-1)p+j_2]} \ ,$$

for any $1 \leq j_1, j_2 \leq p$. For the sake of simplicity, the subscripts (i_1, i_2) are taken modulo p. By definition of $C(\theta)$, it holds that $C_{i_1,i_2} = C_{0,i_2-i_1}$ for any $1 \leq i_1, i_2 \leq p$. Besides, the matrices $C_{0,i}$ are circulant for any $1 \leq i \leq p$. In short, the matrix $C(\theta)$ is of the form

$$C(\theta) = \begin{pmatrix} C_{0,1} & C_{0,2} & \cdots & C_{0,p} \\ \vdots & \vdots & \vdots & \vdots \\ C_{0,p} & C_{0,1} & \cdots & C_{0,p-1} \end{pmatrix},$$

where the matrices $C_{0,i}$ are circulant. Let (i_1, i_2, j_1, j_2) be in $\{1, \ldots, p\}^4$. By definition,

$$C(\theta)_{[(i_1-1)p+j_1,(i_2-1)p+j_2]} = \theta_{[i_2-i_1,j_2-j_1]} \ .$$

Since the matrix θ satisfies condition (3), $\theta_{[i_2-i_1,j_2-j_1]} = \theta_{[i_1-i_2,j_1-j_2]}$. As a consequence, $C(\theta)_{[(i_1-1)p+j_1,(i_2-1)p+j_2]} = C(\theta)_{[(i_2-1)p+j_2,(i_1-1)p+j_1]}$ and $C(\theta)$ is symmetric.

Conversely, let B be a $p^2 \times p^2$ symmetric block circulant matrix. Let us define the matrix θ of size p by

$$\theta_{[i,j]} := B_{[1,(i-1)p+j]} \ ,$$

for any $1 \leq i, j \leq p$. Since the matrix B is block circulant, it follows that $C(\theta) = B$. By definition, $\theta_{[i,j]} = C(\theta)_{[1,(i-1)p+j]}$ and $\theta_{[-i,-j]} = C(\theta)_{[(i-1)p+j,1]}$ for any integers $1 \leq i, j \leq p$. Since the matrix B is symmetric, we conclude that $\theta_{[i,j]} = \theta_{[-i,-j]}$. □

Proof of Lemma 2.2. For any $\theta' \in \Theta^+$, $\gamma_{n,p}(\theta')$ is defined as

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

Applying Lemma 8.15, there exists an orthogonal matrix P that simultaneously diagonalizes Σ and any matrix $C(\theta')$. Let us define $\mathbf{Y}^i := \sqrt{\Sigma}^{-1}\mathbf{X}_i$ and $D_\Sigma := P\Sigma P^*$. Gathering these new notations yields

$$\gamma_{n,p}(\theta') = \frac{1}{p^2} tr\left[\left(I_{p^2} - D(\theta')\right) D_\Sigma \overline{\mathbf{Y}\mathbf{Y}^*} \left(I_{p^2} - D(\theta')\right)\right] \ ,$$

where the vectors $\mathbf{Y}^i$ are independent standard Gaussian random vectors. Except $\overline{\mathbf{Y}\mathbf{Y}^*}$, every matrix involved in this last expression is diagonal. Besides, the diagonal matrix D_Σ is positive since Σ is non-singular. Thus, $tr\left[(I_{p^2} - D(\theta'))D_\Sigma\overline{\mathbf{Y}\mathbf{Y}^*}(I_{p^2} - D(\theta'))\right]$ is almost surely a positive quadratic form on the vector space generated by I_{p^2} and $D(\Theta^+)$. Since the function $D(.)$ is injective and linear on Θ^+, it follows that $\gamma_{n,p}(.)$ is almost surely strictly convex on Θ^+. □

Proof of Lemma 4.1 and Corollary 4.2. The proof only uses the stationarity of the field X on Λ and the l_1 norm of θ. However, the computations are a bit cumbersome. Let θ be an element of Θ^+. By standard Gaussian properties, the expectation of $X_{[0,0]}$ given the remaining covariates is

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_{-\{0,0\}}\right) = \sum_{(i,j)\in\Lambda\backslash(0,0)} \theta_{[i,j]}X_{[i,j]} \ .$$

By assumption $(\mathbb{H}_2)$, the l_1 norm of θ is smaller than one. We shall prove by backward induction that for any subset A of $\Lambda\backslash\{(0,0)\}$ the matrix θ^A uniquely defined by

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_A\right) = \sum_{(i,j)\in A} {\theta^A}_{[i,j]}X_{[i,j]} \text{ and } {\theta^A}_{[i,j]} = 0 \text{ for any } (i,j)\notin A$$

satisfies $\|\theta^A\|_1 \leq \|\theta\|_1$. The property is clearly true if $A = \Lambda\backslash\{(0,0)\}$. Suppose we have proved it for any set of cardinality q larger than one. Let A be a subset of $\Lambda\backslash\{(0,0)\}$ of cardinality $q-1$ and (i,j) be an element of $\Lambda\backslash(A\cup\{(0,0)\})$. Let us derive the expectation of $X_{[0,0]}$ conditionally to X_A from the expectation of $X_{[0,0]}$ conditionally to $X_{A\cup\{(i,j)\}}$.

$$\begin{aligned}\mathbb{E}_\theta\left(X_{[0,0]}|X_A\right) &= \mathbb{E}_\theta\left[\mathbb{E}(X_{[0,0]}|X_A)|X_{A\cup\{(i,j)\}}\right] \\ &= \sum_{(k,l)\in A} {\theta^{A\cup\{(i,j)\}}}_{[k,l]}X_{[k,l]} + {\theta^{A\cup\{(i,j)\}}}_{[i,j]}\mathbb{E}_\theta\left[X_{[i,j]}|X_A\right] \ . \end{aligned} \tag{87}$$

Let us take the conditional expectation of $X_{[i,j]}$ with respect to $X_{A\cup\{(0,0)\}}$. Since the field X is stationary on Λ and by the induction hypothesis, the unique matrix $\theta_{(i,j)}^{A\cup\{(0,0)\}}$ defined by

$$\mathbb{E}_\theta\left(X_{[i,j]}|X_{A\cup\{(0,0)\}}\right) = \sum_{(k,l)\in A\cup\{(0,0)\}} {\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[k,l]}X_{[k,l]}$$

and ${\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[k,l]} = 0$ for any $(k,l)\notin A\cup\{(0,0)\}$ satisfies $\|\theta_{(i,j)}^{A\cup\{(0,0)\}}\|_1 \leq \|\theta\|_1$. Taking the expectation conditionally to X_A of this previous expression leads to

$$\mathbb{E}_\theta\left(X_{[i,j]}|X_A\right) = \sum_{(k,l)\in A} {\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[k,l]}X_{[k,l]} + {\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[0,0]}\mathbb{E}\left(X_{[0,0]}|X_A\right) \ . \tag{88}$$

Gathering identities (87) and (88) yields

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_A\right) = \sum_{(k,l)\in A} \frac{{\theta^{A\cup\{i,j\}}}_{[k,l]} + {\theta^{A\cup\{(i,j)\}}}_{[i,j]}{\theta_{(i,j)}^{A\cup\{0,0\}}}_{[k,l]}}{1-{\theta^{A\cup\{(i,j)\}}}_{[i,j]}{\theta_{(i,j)}^{A\cup\{0,0\}}}_{[0,0]}}X_{[k,l]} \ ,$$

since $\left|\theta^{A\cup\{i,j\}}{}_{[i,j]}\theta^{i,j}_{A\cup\{0,0\}}{}_{[0,0]}\right| < 1$. Then, we upper bound the l_1 norm of θ^A using that $\|\theta^{A\cup\{(i,j)\}}\|_1$ and $\|\theta^{A\cup\{(0,0)\}}_{(i,j)}\|_1$ are smaller or equal to $\|\theta\|_1$.

$$\begin{aligned}
\|\theta^A\|_1 &\leq \frac{1}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|}\left(\sum_{(k,l)\in A}\left|\theta^{A\cup\{j+1\}}{}_{[k,l]}\right| + \sum_{(k,l)\in A}\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{0,0\}}_{(i,j)}{}_{[k,l]}\right|\right)\\
&\leq \frac{\|\theta\|_1 + \left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\left(\sum_{(k,l)\in A\cup\{(0,0)\}}\left|\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[k,l]}\right| - 1 - \left|\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|\right)}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{i,j}{}_{[0,0]}\right|}\\
&\leq \frac{\|\theta\|_1\left(1+\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\right) - \left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\left(1+\left|\theta^{A\cup\{(0,0)\}}_{i,j}{}_{[0,0]}\right|\right)}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|}\\
&\leq \|\theta\|_1 + \frac{\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\left(\|\theta\|_1 - 1\right)\left(1+\left|\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|\right)}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|}\,.
\end{aligned}$$

Since $\|\theta\|_1$ is smaller than one, it follows that $\|\theta^A\|_1 \leq \|\theta\|_1$.

Let m be a model in the collection $\mathcal{M}_1$. Since m stands for a set of neighbors of $(0,0)$, we may define θ^m as above. It follows that $\|\theta^m\|_1 \leq \|\theta\|_1$. Since the field X is stationary on the torus, X follows the same distribution as the field X^s defined by $X^s{}_{[i,j]} = X_{[-i,-j]}$. By uniqueness of θ^m, we obtain that $\theta^m{}_{[i,j]} = \theta^m{}_{[-i,-j]}$. Thus, θ^m belongs to the space Θ_m. Moreover, θ^m minimizes the function $\gamma(.)$ on Θ_m. Since the l_1 norm of θ^m is smaller than one, θ^m belongs to $\Theta^+_{m,2}$. The matrices θ^m and θ_{m,ρ_1} are therefore equal, which concludes the proof in the non-isotropic case.

Let us now turn to the isotropic case. Let θ belong to $\Theta^{\mathrm{iso},+}$ and let m be a model in $\mathcal{M}_1$. As previously, the matrix θ^m satisfies $\|\theta^m\|_1 \leq \|\theta\|_1$. Since the distribution of X is invariant under the action of the group G, θ^m belongs to Θ^{iso}_m. Since $\|\theta^m\|_1 \leq \|\theta\|_1$, θ^m lies in $\Theta^{+,\mathrm{iso}}_{m,2}$. It follows that $\theta^m = \theta^{\mathrm{iso}}_{m,\rho_1}$. □

Proof of Corollary 4.3. Let θ be a matrix in Θ^+ such that $(\mathbb{H}_2)$ holds and let m be a model in $\mathcal{M}_1$. We decompose $\gamma(\widehat{\theta}_{m,\rho_1})$ using the conditional expectation of $X_{[0,0]}$ given X_m.

$$\begin{aligned}
\gamma(\widehat{\theta}_{m,\rho_1}) &= \mathbb{E}_\theta\left[X_{[0,0]} - \sum_{(i,j)\in m}\widehat{\theta}_{m,\rho_1}{}_{[i,j]}X_{[i,j]}\right]^2\\
&= \mathbb{E}_\theta\left[X_{[0,0]} - \mathbb{E}_\theta\left(X_{[0,0]}\,|X_m\right)\right]^2 + \mathbb{E}_\theta\left[\mathbb{E}_\theta\left(X_{[0,0]}\,|X_m\right) - \sum_{(i,j)\in m}\widehat{\theta}_{m,\rho_1}{}_{[i,j]}X_{[i,j]}\right]^2.
\end{aligned}$$

By Corollary 4.2, we know that

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

Combining these two last identities yields

$$\gamma(\widehat{\theta}_{m,\rho_1}) \quad = \quad \gamma(\theta_{m,\rho_1}) + \mathbb{E}_\theta \left[\sum_{(i,j)\in\Lambda\setminus\{(0,0)\}} \left(\theta_{m,\rho_1} - \widehat{\theta}_{m,\rho_1}\right)_{[i,j]} X_{[i,j]} \right]^2 .$$

Subtracting $\gamma(\theta)$, we obtain the first result. The proof is analogous in the isotropic case. □

Acknowledgements

I gratefully thank Pascal Massart for many fruitful discussions.

References

[BBLM05] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005.

[BD91] P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1991.

[Bes75] J. E. Besag. Statistical Analysis of Non-Lattice Data. *The Statistician*, 24(3):179–195, 1975.

[Bes77] J. E. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3):616–618, 1977.

[Bil95] P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[Bir05] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory*, 51(4):1611–1615, 2005.

[BK95] J. E. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995.

[BM75] J. E. Besag and P. A. P. Moran. On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, 62(3):555–562, 1975.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[Cre93] N. A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1993.

[CV08] N. A. C. Cressie and N. Verzelen. Conditional-mean least-squares of Gaussian Markov random fields to Gaussian fields. *Comput. Statist. Data Analysis*, 52(5):2794–2807, 2008.

[Edw00] D. Edwards. *Introduction to graphical modelling*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2000.

[Gra06] R.M. Gray. *Toeplitz and Circulant Matrices: A Review*. Now Publishers, Norwell, Massachusetts, rev. edition, 2006.

[Guy87] X. Guyon. 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)*, volume 11 of *Travaux Rech.*, pages 15–62. Publ. Fac. Univ. Saint-Louis, Brussels, 1987.

[Guy95] X. Guyon. *Random fields on a network*. Probability and its Applications (New York). Springer-Verlag, New York, 1995.

[GY99] X. Guyon and J.F. Yao. On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Multivariate Anal.*, 70(2):221–249, 1999.

[HT89] C. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[KC84] R. Kashyap and R. Chellapa. Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Transactions on Information Theory*, 29:60–72, 1984.

[Lau96] S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[LD93] S. Lakshmanan and H. Derin. Valid parameter space for 2-D Gaussian Markov random fields. *IEEE Trans. Inform. Theory*, 39(2):703–709, 1993.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

[Mor73] P. A. P. Moran. A Gaussian Markovian process on a square lattice. *J. Appl. Probability*, 10:54–62, 1973.

[MT98] A. D. R. McQuarrie and C.-L. Tsai. *Regression and time series model selection*. World Scientific Publishing Co. Inc., River Edge, NJ, 1998.

[RBLZ08] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

[RH05] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London, 2005.

[Ros85] M. Rosenblatt. *Stationary sequences and random fields*. Birkhäuser Boston Inc., Boston, MA, 1985.

[RT02] H. Rue and H. Tjelmeland. Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, 29(1):31–49, 2002.

[SFG08] H.-R. Song, M. Fuentes, and S. Ghosh. A comparative study of gaussian geostatistical models and gaussian markov random field models. *Journal of Multivariate Analysis*, 99:1681–1697, 2008.

[Shi80] R. Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, 8(1):147–164, 1980.

[Tal96] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[Ver09] N. Verzelen. Data-driven neighborhood selection of a gaussian field. Technical Report RR-6798, INRIA, 2009.

[Yu97] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399