THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Statistique*

Présentée et soutenue par *Lilian MUÑIZ ALVAREZ*
Le *19 novembre 2010*

Titre : *Estimation nonparamétrique de la structure de covariance des processus stochastiques*

JURY

M. Alejandro MAASS	*Universidad de Chile*	*Président*
Mme Fabienne COMTE	*Université Paris Descartes*	*Rapporteur*
M. Sébastien VAN BELLEGEM	*Université Toulouse I*	*Rapporteur*
Mme Karin BERTIN	*Universidad de Valparaiso*	*Examinateur*
M. Li-Vang LOZADA CHANG	*Universidad de La Habana*	*Examinateur*

Ecole doctorale : *M.I.T.T.*
Unité de recherche : *Institut de Mathématiques de Toulouse*
Directeur(s) de Thèse :
M. Jérémie BIGOT Université Paul Sabatier
M. Rolando J . BISCAY LIRIO Universidad de Valparaiso
M. Jean-Michel LOUBES Université Paul Sabatier

THÉSE

présentée en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ PAUL SABATIER TOULOUSE III

Discipline : Mathématiques
Spécialité : Statistique

par

Lilian Muñiz Alvarez

Estimation Nonparamétrique de la Structure de Covariance des Processus Stochastiques

Soutenue le 19 Novembre 2010 devant le jury composé de :

M. Alejandro Maass	Universidad de Chile	Président
Mme. Fabienne Comte	Université Paris Descartes	Rapporteur
M. Sébastien Van Bellegem	Université Toulouse I	Rapporteur
Mme. Karin Bertin	Universidad de Valparaiso	Examinateur
M. Li-Vang Lozada Chang	Universidad de La Habana	Examinateur
M. Jérémie Bigot	Université Paul Sabatier	Directeur de thèse
M. Rolando J. Biscay Lirio	Universidad de Valparaiso	Directeur de thèse
M. Jean–Michel Loubes	Université Paul Sabatier	Directeur de thèse

Institut de Mathématiques de Toulouse
UMR CNRS 5219, Université Paul Sabatier, Toulouse III

À ma famille
A mi familia

Remerciements

Je tiens tout d'abord à exprimer ma sincère gratitude à mes directeurs de thèse Jean-Michel Loubes, Jérémie Bigot et Rolando J. Biscay Lirio pour la confiance qu'ils m'ont témoignée en acceptant d'encadrer ce travail, pour leur entière disponibilité, leurs précieuses idées et leurs nombreuses suggestions et remarques qui m'ont été d'un apport bénéfique. J'ai été particulièrement touché par leur gentillesse et tout le soutien qu'ils m'ont apporté pendant ces années passées ensemble.

Merci beaucoup à Fabrice Gamboa et Li-Vang Lozada Chang grâce à qui la collaboration entre l'Université de La Havane et l'Université Paul Sabatier a lieu. Je souhaite aussi remercier à nouveau à mes directeurs de thèse qui, dès mon arrivée en France, m'ont aidé à résoudre les difficultés de toute nature que j'ai rencontrées pendant mes séjours en France.

Je tiens à remercier de manière très spéciale à Egide et au CNRS, grâce auxquels cette recherche a pu être financée. Merci beaucoup aussi aux programmes de coopération MELISSA et PREFALC pour propicier et motiver la collaboration académique avec les pays latino-americains.

Je remercie également Fabienne Comte et Sébastien Van Bellegem pour avoir accepté d'être rapporteurs de cette thèse, pour leurs lectures attentives du manuscript, leurs commentaires me permettant d'envisager des prolongements et l'honneur qu'ils me font en acceptant d'être membre du jury. Merci également à Alejandro Maass, Karin Bertin et Li-Vang Lozada Chang pour avoir accepté de faire parti du jury. J'envoie un remerciement spécial au Centre de Modélisation Mathématique de l'Université du Chili, par son gentil accueil pendant les jours où a eu lieu la soutenance de la thèse.

J'adresse un salut amical aux membres de l'Equipe de Statistique et Probabilités de l'Institut de Mathématique de Toulouse et en particulier à Marie-Laure Ausset pour leur aide dans l'apprentissage de la langue française et avec qui j'ai partagé de très bons moments. Je veux remercier aussi l'ensemble des professeurs de ma faculté à La Havane pour le grand soutien pendant l'achèvement de cette thèse.

Merci beaucoup aussi à tous ceux qui m'ont aidé et m'ont supporté pendant mes séjours en France et qui sont devenus ma famille en France. Particulièrement je voudrais remercier à me grand ami Luis A. Salomón, sur qui je peu compter toujours ; à Ariadna Fuentes, pour l'amitié rapide et sincère qu'elle m'a donné à Toulouse ; à Mary Ana Allen, pour être une très bonne amie ; à Alexander Alvarez, pour m'enseigner les secrets de la vie à Toulouse, à Ricardo Franklin, pour sa joyeuse amitié ; à Angélica, Ignacio, Esteban, David, Jorge, Michael, Santiago et Karoll, pour les moments heureux et merveilleux passés ensemble ; à Ariel, Fanny et Maidelis, pour son amitié inconditionnelle et pour m'aider chaque fois que j'en avais besoin ; à Michele, Edmée, Iván et Lolita, pour me faire parti de leur belle famille ; à Toni, pour l'amitié et l'aide qu'il m'a donné, même avant de me connaître ; à Perrine, pour être la première amie que j'ai rencontré à Toulouse, à Anita, pour les invitations aux matchs de rugby, et à tous les amis que j'ai connus au cours de cette thèse, sans lesquels je ne tiendrai pas tant de beaux souvenirs de mon séjour à Toulouse.

Merci beaucoup aussi à mes amis Yadira, Lissette, Adriana, Carlos, Celia, Wilfredo et Erddys, pour les joies et les peines partagées depuis tant d'années et pour la grande amitié que vous m'offrez tous les jours.

Des remerciements spéciaux à ma mère Lilliam, mon père Germán, mon frère David et mes grandparents Gisela, Camilo, María Beatriz et Ricardo. Merci beaucoup à mon mari bien-aimé, Luis, pour son amour inconditionnel, pour l'immense soutien qu'il m'a toujours donné, pour avoir supporté tous ces mois de séparation, et pour partager avec moi des expériences inoubliables et très romantiques dans les belles villes européennes. Merci beaucoup aussi à Victoria, Luis, Anita et Luisi, pour me faire sentir comme chez moi, pour l'amour et le soutien qu'ils me donnent toujours.

En vue de n'oublier personne, et je suis conscient que je n'ai pas parlé ici de tous ceux qui m'ont aidé d'une façon ou d'une autre, merci beaucoup à vous tous !

Agradecimientos

En primer lugar quisiera expresar mi más sincero agradecimiento a mis tutores Jean-Michel Loubes, Jérémie Bigot y Rolando J. Biscay Lirio, por la confianza que me demostraron al aceptar la supervisión de este trabajo de tesis, por sus valiosas ideas y sugerencias, que me resultaron de gran ayuda, y por ser mi ejemplo a seguir en mi vida profesional. Les agradezco infinitamente su gran amabilidad y todo el apoyo que me han brindado durante estos años en que hemos trabajado juntos.

Muchas gracias a Fabrice Gamboa y Li-Vang Lozada Chang, artífices de la colaboración entre la Universidad de La Habana y la Universidad Paul Sabatier, gracias a los cuales ha sido posible la realización de esta tesis en cotutela entre ambas universidades. También me gustaría agradecer una vez más a mis tutores, que desde mi llegada a Francia, me acogieron y ayudaron grandemente durante mis cuatro estancias en Toulouse.

Quisiera agradecer de manera muy especial a Egide y al CNRS, gracias a los cuales esta investigación ha podido financiarse. Muchas gracias también a los programas de cooperación MELISSA Y PREFALC por propiciar y motivar la colaboración académica con los países latinoamericanos.

También doy muchísimas gracias a Fabienne Comte y Sébastien Van Bellegem por haber aceptado ser los oponentes de esta tesis, por su cuidadosa lectura del manuscrito, sus útiles comentarios y por el honor que me hacen al ser miembros del jurado. Gracias también a Alejandro Maass, Karin Bertin y Li-Vang Lozada Chang por haber aceptado ser parte del jurado. Un agradecimiento especial al Centro de Modelamiento Matemático de la Universidad de Chile, por su amable acogida durante los días en que tuvo lugar la defensa de la tesis.

Agradezco también a todos los miembros del Equipo de Estadística y Probabilidades del Instituto de Matemática de Toulouse, y en especial a Marie-Laure Ausset, por su ayuda en el aprendizaje del idioma francés y con quien he compartido muy buenos momentos. También quiero agradecer a todos mis colegas de la Facultad de Matemática y Computación de la Universidad de La Habana por haberme apoyado mucho durante la realización de esta tesis.

Muchísimas gracias también a todos aquellos que me ayudaron y acompañaron durante mi estancia en Francia, gracias a los cuales me fue muy fácil la adaptación y la convivencia en este bello país, y que se han convertido en mi familia en Francia. Quiero agradecer a Luis A. Salomón, por ser el mejor de los amigos, con el que se puede contar siempre, tanto en lo personal como en lo profesional; a Ariadna Fuentes, por la rápida y sincera amistad que me brindó al conocernos en Toulouse; a Mary Ana Allen, por ser mi pana para siempre, a Alexander Alvarez por enseñarme los secretos de la vida en Toulouse, a Ricardo Franklin, por su alegre amistad; a Angélica, Ignacio, Esteban, David, Jorge, Michael, Santiago y Karoll, por los alegres y maravillosos momentos pasados juntos ; a Ariel, Fanny y Maidelis, por su incondicional amistad y su gran ayuda siempre que la necesité ; a Michele, Edmée, Iván y Lolita, por acogerme en su casa como un miembro más de su bella familia, a Toni, por la amistad y ayuda que me brindó incluso antes de conocerme ; a Perrine, por ser la primera amiga que encontré en Toulouse ; a Anita, por las invitaciones a los partidos de rugby ; en fin, a todos los grandes amigos que he conocido durante la realización de esta tesis, sin los cuales no guardaría tantos bellos recuerdos de mi estancia en Toulouse.

Muchas gracias también a mis amigos Yadira, Lissette, Adriana, Cueto, Celia, Wilfredo y Erddys, por las alegrías y tristezas compartidas durante tantos años y por la inmensa amistad que me brindan cada día.

Un agradecimiento especial a mi madre Lilliam, mi padre German, mi hermano David, y mis abuelos Gisela, Camilo, María Beatriz y Ricardo, por quererme y apoyarme siempre. Muchas gracias a mi adorado esposo Luis, por su amor incondicional, por el apoyo inmenso que me ha dado siempre, por soportar los meses de separación, y por compartir a mi lado paseos inolvidables y muy románticos por bellas ciudades europeas. Muchas gracias también a Victoria, Luis, Anita y Luisi, por hacerme sentir como en mi propia casa, por el cariño y apoyo que me brindan siempre.

Muchas gracias a todos los que me han ayudado durante la realización de esta tesis !

Contents

Notations

If C is a constant, the notation $C(\cdot)$ specifies the dependency of C on some quantities.

w.r.t	with respect to
i.i.d.	independent and identically distributed
n.n.d.	non-negative definite
$\mathbb{N}$	set of all positive integers
$\mathbb{R}$	set of all real numbers
$\mathbb{R}_+$	set of all non-negative real numbers
$\mathbb{R}_+^*$	set of all positive real numbers
$\mathbb{R}^n$	set of all real $n \times 1$ vectors
$\mathbb{R}^{n\times m}$	set of all real $n \times m$ matrices
$\mathbb{P}$	probability measure
$\mathbb{E}$	expectation w.r.t. $\mathbb{P}$
$\mathbb{V}$	variance w.r.t. $\mathbb{P}$
Cov	covariance w.r.t. $\mathbb{P}$
$I\left(\mathcal{A}\right)$	indicator function of the set $\mathcal{A}$
$\lvert\mathcal{A}\rvert$	cardinality of the set $\mathcal{A}$
$\mathcal{A}^c$	complement of the set $\mathcal{A}$
$E_1 \oplus E_2$	direct sum of spaces E_1 and E_2
$\log x$	natural logarithm of $x \in \mathbb{R}_+^*$
$\exp\left(x\right)$	exponential function, the same that e^x, $x \in \mathbb{R}$
$\lvert x\rvert$	absolute value of $x \in \mathbb{R}$
$x \vee y$	maximum of x and y
$x \wedge y$	minimum of x and y
x_+	positive part of x, i.e. $0 \vee x$
$\mathbf{1}_n$	the $n \times 1$ vector with unit elements
$\mathbf{I}_n$	identity matrix of size $n \times n$
$\mathrm{Tr}(\mathbf{A})$	trace of the matrix $\mathbf{A}$
$rk(\mathbf{A})$	rank of the matrix $\mathbf{A}$
$\mathbf{A}^\top$	transpose of the matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse of the matrix $\mathbf{A}$
$\mathbf{A}^-$	pseudo-inverse of the matrix $\mathbf{A}$

$diag\left(\mathbf{A}\right)$	diagonal matrix with same diagonal elements as $\mathbf{A}$
$\inf$	infimum
$\sup$	supremum
$\min$	minimum
$\max$	maximum
$\mathcal{F}$	set of all functions $f : \mathcal{A} \subset \mathbb{R} \to \mathbb{R}$
$\arg\min\limits_{x\in\mathcal{A}} f(x)$	argument of the mimimum of the function $f \in \mathcal{F}$ on the set $\mathcal{A}$
$\lVert f \rVert_{L_2(\mathcal{A})}$	L_2-norm of the function $f \in \mathcal{F}$, i.e. $\lVert f \rVert_{L_2(\mathcal{A})} = \left(\int_{\mathcal{A}} f(x)^2 dx\right)^{\frac{1}{2}}$
$\langle f, g \rangle$	L_2-scalar product of functions $f, g \in \mathcal{F}$
$L_2(\mathcal{A})$	set of functions $f \in \mathcal{F}$ such that $\lVert f \rVert_{L_2(\mathcal{A})} < +\infty$
$span\{f_1, ..., f_k\}$	linear span of the functions $f_1, ..., f_k \in \mathcal{F}$
$\lVert f \rVert_\infty$	uniform norm of the function $f \in \mathcal{F}$, i.e. $\lVert f \rVert_\infty = \sup\limits_{x\in\mathcal{A}} \lvert f\left(x\right) \rvert$
$\lVert \mathbf{x} \rVert_{\ell_p}$	ℓ_p-norm of $\mathbf{x} \in \mathbb{R}^n$, i.e. $\lVert \mathbf{x} \rVert_{\ell_p} = \left(\sum\limits_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$, $p \in \mathbb{N}$
$\langle \mathbf{x}, \mathbf{y} \rangle_{\ell_2}$	ℓ_2-scalar product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
$\lVert \mathbf{x} \rVert_\infty$	uniform norm of $\mathbf{x} \in \mathbb{R}^n$, i.e. $\lVert \mathbf{x} \rVert_\infty = \max\limits_{i=1,\dots,n} \lvert x_i \rvert$
$\rho_{\max}\left(\mathbf{A}\right)$	maximum eigenvalue of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$
$\rho_{\min}\left(\mathbf{A}\right)$	smallest eigenvalue of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$
$\tau\left(\mathbf{A}\right)$	spectral radius of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, i.e. $\tau\left(\mathbf{A}\right)$ is the maximum of the absolute values of the eigenvalues of $\mathbf{A}$
$\lVert \mathbf{A} \rVert_2$	Operator norm of matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, i.e. $\lVert \mathbf{A} \rVert_2 = \sup\limits_{\mathbf{x}\in\mathbb{R}^m : \mathbf{x}\neq 0} \frac{\lVert \mathbf{A}\mathbf{x} \rVert_{\ell_2}}{\lVert \mathbf{x} \rVert_{\ell_2}}$
$\lVert \mathbf{A} \rVert_F$	Frobenious norm of matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, i.e. $\lVert \mathbf{A} \rVert_F = \sqrt{\mathrm{Tr}\left(\mathbf{A}^\top \mathbf{A}\right)}$
$\langle \mathbf{A}, \mathbf{B} \rangle_F$	Frobenious scalar product of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$
$vec\left(\mathbf{A}\right)$	vectorization of matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ (Definition A.4)
$\mathcal{S}_n$	linear espace of $n \times n$ symmetric matrices
$vech\left(\mathbf{A}\right)$	half-vectorization of matrix $\mathbf{A} \in \mathcal{S}_n$ (Definition A.5)
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$ (Definition A.6)
$\square$	end of a proof
$\to$	tends to

Résumé

Introduction

L'objectif principal de cette thèse est de développer des méthodes nonparamétriques pour l'estimation de la covariance d'un processus stochastique. En supposant des conditions différentes sur le processus, des estimateurs de la fonction de covariance seront introduits, possédant la propriété d'être des fonctions définies non-négatives. En outre, une méthode pour l'estimation de la matrice de covariance d'un processus stochastique dans un cadre de grande dimension sera proposé.

L'estimation de la covariance est un problème fondamental dans l'inférence des processus stochastiques, avec de nombreuses applications, allant de la géologie, la météorologie, les séries financières, l'épidémiologie, etc. Il se présente à chaque fois qu'il faut obtenir une bonne prédiction à partir d'une séquence d'observations.

Les méthodes paramétriques ont été largement étudiées dans la littérature statistique (voir Cressie [28] pour une revue), mais ce sont les procédures nonparamétriques qui ont reçu récemment une attention croissante (voir Shapiro et Botha [85], Sampson et Guttorp [80], Hall et al. [47], Perrin et al. [75], Guillot et al. [46], Elogne, Perrin et Thomas-Agnan [40]). Une des difficultés principales des méthodes nonparamétriques est l'obtention d'estimateurs de la fonction de covariance qui soient également des fonctions définies non-négatives. C'est dans ce cadre que nous nous sommes proposés dans cette thèse de construire de nouveaux estimateurs satisfaisant cette propriété.

L'estimation des matrices de covariance dans un cadre de grande dimension, en particulier dans les situations où la dimension des données est comparable ou supérieure à la taille de l'échantillon, a attiré beaucoup d'attention récemment. L'abondance de données de grande dimension est une des raisons de l'intérêt pour ce problème. Une autre raison est l'omniprésence de la matrice de covariance dans les outils d'analyse de données. Par exemple l'Analyse des Composantes Principales (ACP) exige une estimation de la matrice de covariance. Enfin, les progrès récents de la théorie des matrices aléatoires (voir Johnstone [49] et Paul [74] pour une revue) ont permis des études théoriques de l'estimateur empirique de la matrice de covariance, et ont montré que sans régularisation, la covariance empirique n'a pas un bon comportement dans un cadre de grande dimension. Ces résultats ont contribué à motiver la recherche d'autres estimateurs de la matrice de covariance. Dans cette thèse, nous proposons une méthode d'estimation pénalisée pour la matrice de covariance d'un processus stochastique dans un cadre de grande dimension.

Nous avons choisi d'organiser la synthèse de nos travaux sous la forme suivante : une première partie d'introduction générale, le Chapitre 1, où nous présentons de façon succincte les notions à la base de notre travail ainsi que les définitions des différents objets qui nous intéressent. Ensuite, viennent trois chapitres où sont détaillées les nouvelles méthodes d'estimation proposées. Plus précisement, dans chaque chapitre nous avons developpé des techniques d'estimation nonparamétrique différentes : approximation des fonctions par ondelettes avec seuillage dans le Chapitre 2, sélection des modèles dans le Chapitre 3, et estimation par une méthode de pénalisation de type Group-Lasso dans le Chapitre 4.

L'estimation de la fonction de covariance est fortement liée à l'estimation de la densité spectrale du processus. C'est pourquoi au Chapitre 2 nous étudions le problème de l'estimation nonparamétrique de la densité spectrale d'un processus Gaussien stationnaire. A cet effet, nous considérons une méthode qui combine les idées d'estimation des fonctions par méthodes de projection sur une base d'ondelettes avec seuillage et l'estimation par projection de l'information, en vue d'assurer que l'estimateur obtenu a la propriété d'être une fonction non-négative. La densité spectrale du processus est estimée par projection, sur une famille de fonctions exponentielles, de l'approximation par ondelettes avec seuillage du périodogramme. Cela permet de garantir que l'estimateur de la densité spectrale soit une fonction strictement positive. Puis, par le théorème de Bochner, nous obtenons un estimateur défini non-négatif de la fonction de covariance. Le comportement théorique de l'estimateur est établi en termes de la vitesse de convergence de la divergence de Kullback-Leibler sur des classes de Besov. Nous montrons également la bonne performance pratique de l'estimateur dans quelques exemples numériques.

Dans le Chapitre 3, nous proposons une approche par sélection de modèle pour l'estimation de la covariance d'un processus stochastique sous des hypothèses très générales pour le processus. En particulier, nous ne supposons pas que le processus est Gaussien ou stationnaire. En observant des réplications indépendantes et identiquement distribuées du processus à des points d'observation fixes, nous construisons un estimateur de la fonction de covariance en utilisant une approximation du processus par une collection de fonctions. Nous étudions les propriétés non asymptotiques de cet estimateur et nous proposons une façon pour choisir le meilleur estimateur parmi un ensemble de candidats possibles. L'optimalité de la procédure est prouvée par une inégalité oracle qui garantit que le meilleur modèle est sélectionné. Des exemples numériques montrent la bonne performance de l'estimateur proposé.

Enfin, au Chapitre 4, nous présentons l'estimateur Group-Lasso de la matrice de covariance d'un processus stochastique. Nous proposons d'estimer la matrice de covariance dans un cadre de grande dimension sous l'hypothèse que le processus possède une représentation sparse dans un dictionnaire de fonctions de base. En utilisant un modèle de régression matriciel, nous proposons une nouvelle méthodologie pour l'estimation de la matrice de covariance en grande dimension basée sur la régularisation d'un contraste empirique par une pénalité de type Group-Lasso. En utilisant cette pénalité, la méthode sélectionne un ensemble sparse de fonctions de base dans le dictionnaire utilisé pour approcher le processus, en conduisant à une approximation de la matrice de covariance dans un espace de plus faible dimension. En conséquence, nous proposons une nouvelle méthode de réduction de la dimension de données en grande dimension, et étudions son comportement dans la pratique.

Estimation adaptative des fonctions de covariance par seuillage et projection d'information

Le théorème de Bochner affirme qu'une fonction continue sur $\mathbb{R}^d$ est définie non-négatif si et seulement si elle est la transformée de Fourier d'une mesure bornée non-négatif appelée la mesure spectrale. Quand une mesure spectrale a une densité, cette densité est appelée la densité spectrale. Ainsi, l'obtention d'un estimateur défini non-négatif de la fonction de covariance est fortement liée à l'obtention d'un estimateur non-négatif de la densité spectrale du processus. C'est l'idée fondamentale utilisée dans ce chapitre.

L'inférence dans le domaine spectral utilise le périodogramme des données. Le périodogramme n'est pas un bon estimateur de la densité spectrale, il doit être modifié afin d'assurer la consistance. Pour des densités spectrales très régulières, des techniques de lissage linéaire, comme les méthodes de lissage par noyaux sont appropriées (voir Brillinger [23]). Toutefois, les méthodes linéaires ne sont pas capables d'atteindre la vitesse optimale pour des densités spectrales dont la regularité est distribuée de façon non homogène sur le domaine d'intérêt. Pour cela des méthodes non linéaires sont nécessaires. Une méthode non linéaire pour l'estimation adaptative de la densité spectrale d'une séquence Gaussien stationnaire a été proposé par Comte [27]. Elle est basée sur des techniques de sélection de modèles. D'autres procédures non linéaires de lissage sont obtenues en seuillant les coefficients dans une base d'ondelettes, d'abord proposée par Donoho et Johnstone [35]. Dans ce contexte, des règles de seuillage différentes ont été proposées par Neumann [72] et par Fryzlewics, Nason et von Sachs [42] pour n'en citer que quelques-uns.

L'approche de Neumann [72] consiste à pré-estimer la variance du périodogramme par un lissage par noyau, auquel est ensuite appliquée une procédure d'estimation par ondelettes. L'estimation par noyau peut ne pas être appropriée dans les cas où la densité spectrale est de faible régularité. Une façon d'éviter ce problème est proposée dans Fryzlewics, Nason et von Sachs [42], où les seuils des coefficients d'ondelette empiriques sont construits par des pondérations locales appropriées de la norme ℓ_1 du périodogramme. Leurs méthodes ne produisent pas des estimateurs non-négatifs de la densité spectrale, et donc les estimateurs correspondants de la fonction de covariance ne possèdent pas la propriété d'être définis non-négatifs.

Pour pallier ces inconvénients, nous proposons dans ce chapitre, une nouvelle méthode basée sur des estimateurs par ondelettes pour l'estimation de la densité spectrale d'un processus Gaussien stationnaire et de sa fonction de covariance. Comme solution pour assurer la non-negativité de l'estimateur de la densité spectrale, notre méthode associe les idées de seuillage par ondelettes et l'estimation par projection d'information. Nous estimons la densité spectrale par une projection de l'approximation non linéaire par ondelettes du périodogramme sur une famille de fonctions exponentielles. Par conséquent, l'estimateur est non-négatif par construction. Puis, par le théorème de Bochner, l'estimateur correspondant de la fonction de covariance satisfait la propriété d'être défini non-négatif. Cette technique a été étudiée par Barron et Sheu [9] pour l'approximation des fonctions de densité par des séquences de familles exponentielles, par Loubes et Yan [60] pour l'estimation pénalisée de l'estimateur de maximum vraisemblance avec une pénalité ℓ_1, par Antoniadis et Bigot [3] pour l'étude des problèmes inverses de Poisson, et par Bigot et Van Bellegem

[15] pour la déconvolution de la log-densité.

L'optimalité théorique des estimateurs de la densité spectrale d'un processus stationnaire est généralement étudiée en utilisant des bornes de risque pour la perte L_2. C'est le cas dans les articles de Neumann [72], Comte [27] et Fryzlewics, Nason et von Sachs [42]. Dans ce chapitre, le comportement de l'estimateur proposé est établi en fonction de la vitesse de convergence de la divergence de Kullback-Leibler sur des classes de Besov, ce qui est peut-être une fonction de perte plus naturelle pour l'estimation de la densité spectrale que la L_2-norme. De plus, les règles de seuillage que nous utilisons pour obtenir les estimateurs adaptatifs sont différentes des approches antérieures basées sur la décomposition dans des bases d'ondelettes et sont très simples à calculer.

Cadre statistique

Nous considérons que la séquence $(X_t)_{t\in\mathbb{N}}$ satisfait les conditions suivantes :

Hypothèse 1. *La séquence $(X_1, ..., X_n)$ est un échantillon de taille n obtenue à partir d'une suite stationnaire de variables aléatoires Gaussiennes.*

Soit σ la fonction de covariance du processus, $\sigma(h) = Cov(X_t, X_{t+h})$ avec $h \in \mathbb{Z}$. La densité spectrale f est définie comme :

$$f(\omega) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \sigma(h) e^{-i2\pi\omega h},\ \omega \in [0,1].$$

Nous avons besoin de l'hypothèse suivante sur σ :

Hypothèse 2. *La fonction de covariance σ est définie non-negative, et il existe deux constantes $0 < C_1, C_2 < +\infty$ telles que $\sum\limits_{h\in\mathbb{Z}} |\sigma(h)| = C_1$ et $\sum\limits_{h\in\mathbb{Z}} |h\sigma^2(h)| = C_2$.*

Les données sont constituées d'un certain nombre d'observations $X_1, ..., X_n$ en des points régulièrement espacés. Pour assurer la non-négativité de notre estimateur, nous allons chercher des approximations sur une famille exponentielle. Pour cela, nous construisons un ensemble de fonctions exponentielles définies sur une base d'ondelettes.

Soient $\phi(\omega)$ et $\psi(\omega)$, respectivement, la fonction d'échelle et la fonction d'ondelette générées par une décomposition multirésolution orthonormée de $L_2([0,1])$, voir Mallat [63] pour un exposé détaillé sur l'analyse multirésolution en ondelettes. Dans le présent document, les fonctions ϕ et ψ ont des supports compacts et sont telles que $\|\phi\|_\infty < +\infty$ et $\|\psi\|_\infty < +\infty$. Pour tout entier $j_0 \geq 0$, toute fonction $g \in L_2([0,1])$ a la représentation suivante :

$$g(\omega) = \sum_{k=0}^{2^{j_0}-1} \langle g, \phi_{j_0,k} \rangle \phi_{j_0,k}(\omega) + \sum_{j=j_0}^{+\infty} \sum_{k=0}^{2^j-1} \langle g, \psi_{j,k} \rangle \psi_{j,k}(\omega),$$

où

$$\phi_{j_0,k}(\omega) = 2^{\frac{j_0}{2}} \phi\left(2^{j_0}\omega - k\right) \text{ et } \psi_{j,k}(\omega) = 2^{\frac{j}{2}} \psi\left(2^j\omega - k\right).$$

L'idée principale de ce chapitre est d'approximer la densité spectrale f sur cette base d'ondelettes et de trouver un estimateur de cette approximation, qui est ensuite modifiée pour imposer la propriété de positivité. Les coefficients d'échelle et les coefficients d'ondelettes de la densité spectrale f sont définis par

$$a_{j_0,k} = \langle f, \phi_{j_0,k} \rangle = \int_0^1 f(\omega)\, \phi_{j_0,k}(\omega)\, d\omega$$

et

$$b_{j,k} = \langle f, \psi_{j,k} \rangle = \int_0^1 f(\omega)\, \psi_{j,k}(\omega)\, d\omega.$$

Pour simplifier les notations, nous écrivons $(\psi_{j,k})_{j=j_0-1}$ pour les fonctions d'échelle $(\phi_{j,k})_{j=j_0}$. Soit $j_1 \geq j_0$, nous notons Λ_{j_1} l'ensemble

$$\Lambda_{j_1} = \left\{ (j,k) : j_0 - 1 \leq j < j_1, 0 \leq k \leq 2^j - 1 \right\}.$$

Notez que $|\Lambda_{j_1}| = 2^{j_1}$, où $|\Lambda_{j_1}|$ désigne le cardinal de Λ_{j_1}. Soit θ un vecteur dans $\mathbb{R}^{|\Lambda_{j_1}|}$, la famille $\mathfrak{E}_{j_1}$ de fonctions exponentielles à l'échelle j_1 est définie comme l'ensemble des fonctions :

$$\mathfrak{E}_{j_1} = \left\{ f_{j_1,\theta}(.) = \exp\left(\sum_{(j,k)\in\Lambda_{j_1}} \theta_{j,k}\psi_{j,k}(.) \right),\ \theta = (\theta_{j,k})_{(j,k)\in\Lambda_{j_1}} \in \mathbb{R}^{|\Lambda_{j_1}|} \right\}.$$

Nous allons demander que notre estimateur de la densité spectrale appartienne à la famille $\mathfrak{E}_{j_1}$ des fonctions exponentielles, qui sont positives par définition.

D'après Csiszár [29], il est possible de définir la projection d'une fonction f sur $\mathfrak{E}_{j_1}$. Si cette projection existe, elle est définie comme la fonction $f_{j_1,\theta^*_{j_1}}$ dans la famille exponentielle $\mathfrak{E}_{j_1}$ qui est la plus proche de la vraie fonction f dans le sens de la divergence de Kullback-Leibler. Elle est caractérisée comme la seule fonction dans la famille $\mathfrak{E}_{j_1}$ pour laquelle

$$\left\langle f_{j_1,\theta^*_{j_1}}, \psi_{j,k} \right\rangle = \langle f, \psi_{j,k} \rangle := \beta_{j,k} \text{ pour tout } (j,k) \in \Lambda_{j_1}.$$

Notez que la notation $\beta_{j,k}$ est utilisée pour, à la fois, les coefficients d'échelle $a_{j_0,k}$ et les coefficients d'ondelettes $b_{j,k}$.

Soit

$$I_n(\omega) = \frac{1}{2\pi n} \sum_{t=1}^{n} \sum_{t'=1}^{n} \left(X_t - \overline{X}\right) \left(X_{t'} - \overline{X}\right)^* e^{-i2\pi\omega(t-t')},$$

le périodogramme classique, où $\left(X_t - \overline{X}\right)^*$ désigne la transposée conjuguée de $\left(X_t - \overline{X}\right)$ et $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$. La décomposition de $I_n(\omega)$ sur la base d'ondelettes permet d'obtenir des estimateurs de $a_{j_0,k}$ et $b_{j,k}$ donnés par

$$\widehat{a}_{j_0,k} = \int_0^1 I_n(\omega)\, \phi_{j_0,k}(\omega)\, d\omega \quad \text{et} \quad \widehat{b}_{j,k} = \int_0^1 I_n(\omega)\, \psi_{j,k}(\omega)\, d\omega.$$

Il semble donc naturel d'estimer la fonction f en recherchant $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$ tel que

$$\left\langle f_{j_1,\widehat{\theta}_n}, \psi_{j,k} \right\rangle = \int_0^1 I_n(\omega)\,\psi_{j,k}(\omega)\,d\omega := \widehat{\beta}_{j,k} \text{ pour tout } (j,k) \in \Lambda_{j_1}, \tag{1}$$

où $\widehat{\beta}_{j,k}$ désigne l'estimation des coefficients d'échelle $\widehat{a}_{j_0,k}$ et les coefficients d'ondelettes $\widehat{b}_{j,k}$. La fonction $f_{j_1,\widehat{\theta}_n}$ est l'estimateur de projection positif de la densité spectrale.

De même, l'estimateur non linéaire positif avec seuillage dur est défini comme la fonction $f^{HT}_{j_1,\widehat{\theta}_n,\xi}$ (avec $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$) telle que

$$\left\langle f^{HT}_{j_1,\widehat{\theta}_n,\xi}, \psi_{j,k} \right\rangle = \delta_\xi\left(\widehat{\beta}_{j,k}\right) \text{pour tout } (j,k) \in \Lambda_{j_1}, \tag{2}$$

où δ_ξ est la règle de seuillage dur définie par $\delta_\xi(x) = xI(|x| \geq \xi)$ pour $x \in \mathbb{R}$, et $\xi > 0$ est un seuil approprié.

L'existence de tels estimateurs n'est pas garantie a priori. En outre, il n'existe aucun moyen d'obtenir une expression explicite pour $\widehat{\theta}_n$. Dans nos simulations, nous utilisons une approximation numérique de $\widehat{\theta}_n$ qui est obtenue via un algorithme d'optimisation approprié. Prouver que les estimateurs existent avec une probabilité 1 est une tâche difficile. Pour le problème de l'estimation d'une densité à partir d'une suite de variables aléatoires indépendantes et identiquement distribuées, il est même indiqué dans Barron et Sheu [9] que pour certaines familles exponentielles, le vecteur $\widehat{\theta}_n$ peut ne pas exister avec une faible probabilité. Ainsi, dans les sections suivantes, des conditions suffisantes sont données pour l'existence de $f_{j_1,\widehat{\theta}_n}$ et $f^{HT}_{j_1,\widehat{\theta}_n,\xi}$ avec une probabilité qui tend vers 1 quand $n \to +\infty$.

Pour évaluer la qualité des estimateurs, nous allons mesurer la différence entre un estimateur $\widehat{f}$ et la vraie fonction f au sens de l'entropie relative (divergence de Kullback-Leibler) définie par :

$$\Delta\left(f;\widehat{f}\right) = \int_0^1 \left(f \log\left(\frac{f}{\widehat{f}}\right) - f + \widehat{f}\right) d\mu,$$

où μ désigne la mesure de Lebesgue sur $[0,1]$. Il peut être démontré que $\Delta\left(f;\widehat{f}\right)$ est non-negative et égale à zéro si et seulement si $\widehat{f} = f$.

Il est bien connu que les espaces de Besov pour les fonctions périodiques dans $L_2([0,1])$ peuvent être caractérisés en terme de coefficients d'ondelettes (voir par exemple Mallat [63]). Supposons que ψ a m moments nuls, et que $0 < s < m$ est le paramètre de régularité usuel. Alors, pour une boule de Besov $B^s_{p,q}(A)$ de rayon $A > 0$ avec $1 \leq p,q \leq \infty$, on a que pour $s^* = s + 1/2 - 1/p \geq 0$:

$$B^s_{p,q}(A) := \left\{g \in L_2([0,1]) : \|g\|_{s,p,q} \leq A\right\},$$

où

$$\|g\|_{s,p,q} := \left(\sum_{k=0}^{2^{j_0}-1} |\langle g, \phi_{j_0,k}\rangle|^p\right)^{\frac{1}{p}} + \left(\sum_{j=j_0}^{\infty} 2^{js^*q}\left(\sum_{k=0}^{2^j-1} |\langle g, \psi_{j,k}\rangle|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}},$$

avec les sommes précédentes remplacées par un supremum si $p = \infty$ ou $q = \infty$.

La condition $s + 1/2 - 1/p \geq 0$ est imposée pour s'assurer que $B^s_{p,q}(A)$ est un sous-espace de $L_2([0,1])$, et nous nous limiterons à ce cas dans le présent document (bien que ce ne soit pas toujours indiqué, il est clair que tous nos résultats sont valables pour $s < m$).

Soit $M > 0$, nous définissons l'ensemble des fonctions $F^s_{p,q}(M)$ par

$$F^s_{p,q}(M) = \{f = \exp(g) : \ \|g\|_{s,p,q} \leq M\},$$

où $\|g\|_{s,p,q}$ est la norme dans l'espace de Besov $B^s_{p,q}$. Notez que supposer que $f \in F^s_{p,q}(M)$ implique que f est strictement positive. Dans la prochaine section nous établissons la vitesse de convergence de nos estimateurs en fonction de la divergence de Kullback-Leibler sur les classes de Besov.

Comportement asymptotique des estimateurs

Nous faisons l'hypothèse suivante sur la base d'ondelettes qui garantit que l'Hypothèse 2 est vraie uniformément sur $F^s_{p,q}(M)$.

Hypothèse 3. *Soit $M > 0$, $1 \leq p \leq 2$ et $s > \frac{1}{p}$. Pour $f \in F^s_{p,q}(M)$ et $h \in Z$, soit $\sigma(h) = \int\limits_0^1 f(\omega) e^{i2\pi\omega h} d\omega$, $C_1(f) = \sum\limits_{h\in\mathbb{Z}} |\sigma(h)|$ et $C_2(f) = \sum\limits_{h\in\mathbb{Z}} |h\sigma^2(h)|$. Alors, la base d'ondelettes est telle qu'il existe une constante M_* telle que pour tout $f \in F^s_{p,q}(M)$, $C_1(f) \leq M_*$ et $C_2(f) \leq M_*$.*

Estimation par projection

Le théorème suivant est le résultat général sur l'estimateur de projection de la fonction de densité spectrale donné par (1). Notez que le choix du niveau de résolution j_0 est de peu d'importance, et sans perte de généralité nous prenons $j_0 = 0$ pour l'estimateur $f_{j_1,\widehat{\theta}_n}$.

Théorème 1. *Supposons que $f \in F^s_{2,2}(M)$ avec $s > \frac{1}{2}$ et supposons que les Hypothèses 1, 2 et 3 sont satisfaites. Soit $j_1 = j_1(n)$ le plus grand entier tel que $2^{j_1} \leq n^{\frac{1}{2s+1}}$. Alors, avec une probabilité qui tend vers 1 quand $n \to +\infty$, l'estimateur de projection d'information (1) existe et satisfait :*

$$\Delta\left(f; f_{j_1(n),\widehat{\theta}_n}\right) = \mathcal{O}_p\left(n^{-\frac{2s}{2s+1}}\right).$$

De plus, la convergence est uniforme sur toute la classe $F^s_{2,2}(M)$ dans le sens que

$$\lim_{K\to+\infty} \lim_{n\to+\infty} \sup_{f\in F^s_{2,2}(M)} \mathbb{P}\left(n^{\frac{2s}{2s+1}} \Delta\left(f; f_{j_1(n),\widehat{\theta}_n}\right) > K\right) = 0.$$

Ce théorème prouve l'existence avec une probabilité qui tend vers 1 d'un estimateur de la densité spectrale f donnée par $f_{j_1(n),\widehat{\theta}_{j_1(n)}}$. Cet estimateur est strictement positif par construction. Par conséquent, l'estimateur correspondant de la fonction de covariance

$\widehat{\sigma}^L$ (qui est obtenu comme la transformée de Fourier inverse de $f_{j_1(n),\widehat{\theta}_n}$) est une fonction définie positive par le théorème de Bochner. Ainsi $\widehat{\sigma}^L$ est une fonction de covariance.

Dans le problème de l'estimation de densité à partir d'un échantillon i.i.d., Koo [55] a montré que, pour la divergence de Kullback-Leibler, $n^{-\frac{2s}{2s+1}}$ est le taux de convergence optimal pour le problème de l'estimation d'une densité f telle que $\log(f)$ appartient à l'espace $B^s_{2,2}(M)$. Pour des densités spectrales appartenant à une boule de Besov $B^s_{p,q}(M)$ générale, Neumann [72] a également montré que c'est la vitesse de convergence optimale pour le risque L_2. Pour la divergence de Kullback-Leibler, nous conjecturons que $n^{-\frac{2s}{2s+1}}$ est également le taux minimax de convergence pour les densités spectrales appartenant à $F^s_{2,2}(M)$.

Le résultat obtenu dans le théorème ci-dessus est non adaptatif, dans la mesure où la sélection de $j_1(n)$ dépend de la régularité s de f, qui est inconnue. En outre, le résultat n'est vrai que pour des fonctions lisses (comme $F^s_{2,2}(M)$ qui correspond à un espace de Sobolev d'ordre s) et ne permet pas d'atteindre un taux de convergence optimal lorsque, par exemple $g = \log(f)$ a des singularités. Nous proposons donc dans la section suivante un estimateur adaptatif obtenu en appliquant une procédure non linéaire appropriée.

Estimation adaptative

Pour l'estimation adaptative, nous avons besoin de définir une règle de seuillage appropriée pour les coefficients d'ondelettes du périodogramme. Ce seuil dépend du niveau de résolution et dans le présent document prendra la forme :

$$\widehat{\xi}_{j,n} = 2\left[2\left\|\widehat{f}_n\right\|_\infty\left(\sqrt{\frac{\delta}{(1-b)^2}\frac{\log n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty\frac{\delta}{(1-b)^2}\frac{\log n}{n}\right) + \sqrt{\frac{\log n}{n}}\right], \qquad (3)$$

où $\left\|\widehat{f}_n\right\|_\infty$ est un estimateur approprié de $\|f\|_\infty$, $\delta \geq 0$ est un paramètre de réglage et $b \in \left[\frac{3}{4}, 1\right)$. Alors, le théorème suivant est vrai :

Théorème 2. *Supposons que $f \in F^s_{p,q}(M)$ avec $s > \frac{1}{2} + \frac{1}{p}$ et $1 \leq p \leq 2$. Supposons également que les Hypothèses 1, 2 et 3 sont satisfaites. Pour tout $n > 1$, soit $j_0 = j_0(n)$ le plus grand entier tel que $2^{j_0} \geq \log n \geq 2^{j_0-1}$, et soit $j_1 = j_1(n)$ le plus grand entier tel que $2^{j_1} \geq \frac{n}{\log n} \geq 2^{j_1-1}$. Prenez les constantes $\delta = 6$ et $b \in \left[\frac{3}{4}, 1\right)$, et soit $\widehat{\xi}_{j,n}$ le seuil (3). Alors, sous certaines conditions de régularité sur f, l'estimateur de seuillage (2) existe avec une probabilité qui converge vers 1 quand $n \to +\infty$ et satisfait*

$$\Delta\left(f; f^{HT}_{j_0(n),j_1(n),\widehat{\theta}_n,\widehat{\xi}_{j,n}}\right) = \mathcal{O}_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).$$

Notons que nous obtenons finalement non seulement un estimateur entièrement explicite de f qui atteint le taux de convergence optimal sans connaissance préalable de la régularité de la densité spectrale, mais également un estimateur qui est une vraie fonction de covariance.

Estimation nonparamétrique de fonctions de covariance par sélection de modèle

Dans ce chapitre, nous proposons d'utiliser une procédure de sélection de modèle pour construire un estimateur nonparamétrique de la fonction de covariance d'un processus stochastique sous des hypothèses générales pour le processus. En particulier, nous ne supposons ni la Gaussianité ni la stationnarité du processus observé.

Nous considérons un processus stochastique $X(t)$ à valeurs dans $\mathbb{R}$, indexé par $t \in T$, où T est un sous-ensemble de $\mathbb{R}^d$, $d \in \mathbb{N}$. Dans notre travail, nous supposons que sa fonction de covariance est finie, i.e. $|\sigma(s,t)| = |Cov(X(s), X(t))| < +\infty$ pour tout $s, t \in T$, et par simplicité, qu'il a une moyenne nulle $\mathbb{E}(X(t)) = 0$ pour tout $t \in T$. Les observations sont $X_i(t_j)$ pour $i = 1, ..., N$, $j = 1, ..., n$, où les points d'observation $t_1, ..., t_n \in T$ sont fixes, et $X_1, ..., X_N$ sont des copies indépendantes du processus X.

Des approximations fonctionnelles des processus $X_1, ..., X_N$ à partir de données $(X_i(t_j))$ sont utilisées dans l'estimation de la fonction de covariance. Lorsqu'il s'agit d'analyse de données fonctionnelles, le lissage des processus $X_1, ..., X_N$ est parfois réalisée dans un premier temps avant de calculer la covariance empirique telles que l'interpolation spline par exemple (voir Elogne, Perrin and Thomas-Agnan [40]) ou une projection sur une base générique finie. Soit $\mathbf{x}_i = (X_i(t_1), ..., X_i(t_n))^\top$ le vecteur des observations dans les points $t_1, ..., t_n$ avec $i \in \{1, ..., N\}$ et $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\right) = (\sigma(t_j, t_k))_{1 \le j \le n, 1 \le k \le n}$ la matrice de covariance dans les points $t_1, ..., t_n$. Soit $\{g_\lambda\}_\lambda$ une collection de fonctions indépendantes $g_\lambda : T \to \mathbb{R}$, et $\mathcal{M}$ un ensemble dénombrable définie par $\mathcal{M} = \{m : m \text{ est un ensemble d'indices}\}$. Soit $m \in \mathcal{M}$ un sous-ensemble d'indices de taille $|m| \in \mathbb{N}$, nous définissons la matrice $\mathbf{G}$ ayant pour entrées $g_{j\lambda} = g_\lambda(t_j)$, $j = 1, ..., n$, $\lambda \in m$.

Nous considérons que l'estimateur $\widehat{\mathbf{\Sigma}}$ de $\mathbf{\Sigma}$ est donné par $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$, où $\widehat{\mathbf{\Psi}}$ est l'estimateur des moindres carrés du modèle de régression matriciel suivant

$$\mathbf{x}_i\mathbf{x}_i^\top = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, ..., N, \tag{4}$$

où $\mathbf{\Psi}$ est une matrice symétrique et $\mathbf{U}_i$ sont des erreurs matricielles i.i.d. L'une des principales contributions de ce chapitre est de montrer que le modèle (4) permet de traiter une grande variété de cas, et de construire un estimateur optimal de la covariance par sélection de modèle sans hypothèses trop fortes sur le modèle. D'autre part, il sera montré que le modèle (4) conduit à un estimateur $\widehat{\mathbf{\Psi}}$ qui se trouve dans la classe des matrices définies non-négatives, et donne donc une bonne matrice de covariance $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$.

Une méthode similaire a été développée pour l'interpolation lisse de fonctions de covariance dans Biscay et al. [19], mais elle est limitée aux fonctions de base qui sont déterminées par noyaux reproduisants dans des espaces de Hilbert appropriés. Des idées similaires sont également abordées dans Matsuo et al. [67]. Ces auteurs traitent du problème de l'estimation de $\mathbf{\Sigma}$ dans la classe de covariance $\mathbf{\Gamma} = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$ induite par une projection sur une base d'ondelettes orthogonales. Leur fonction de contraste n'est pas générale, car ils choisissent la loi Gaussienne, et donc leur méthode nécessite des hypothèses de distribution. Nous rappelons également que le calcul de la probabilité Gaussienne nécessite l'inversion de $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$, qui n'est pas directement possible si $rk(\mathbf{G}) < n$ ou si certains éléments de la diagonale de la matrice définie non-négative $\mathbf{\Psi}$ sont nuls.

À notre connaissance, aucun travail antérieur n'a proposé d'utiliser le modèle de régression matriciel (4) sous des hypothèses générales sur les moments du processus X en utilisant une approximation par projection sur une base générale, pour l'estimation non paramétrique de la fonction de covariance.

Cadre statistique

Nous supposons que X a des moments finis jusqu'à l'ordre 4. Soit $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top$ la matrice de covariance empirique (non corrigée par la moyenne) des données $\mathbf{x}_1, ..., \mathbf{x}_N$. Notre objectif est de construire un estimateur par sélection de modèles de la covariance du processus X observé avec N répétitions indépendantes. Les asymptotiques seront prises par rapport à N, le nombre de copies du processus.

L'approche que nous allons développer pour estimer la fonction de covariance σ est basée sur les deux ingrédients principaux suivants : d'abord, nous considérons une approximation fonctionnelle $\widetilde{X}$ pour approcher le processus X et dès lors considérer la covariance de $\widetilde{X}$ comme une approximation de la vraie covariance σ.

Pour cela, soit $m \in \mathcal{M}$, nous considérons une approximation du processus X de la forme suivante :

$$\widetilde{X}\left(t\right) = \sum_{\lambda \in m} a_\lambda g_\lambda\left(t\right), \tag{5}$$

où a_λ sont des coefficients aléatoires appropriés. Par conséquent, il est naturel de considérer la fonction de covariance ρ de $\widetilde{X}$ comme une approximation de σ. La covariance ρ peut être écrite comme

$$\rho\left(s,t\right) = \mathbf{G}_s^\top \overline{\mathbf{\Psi}}\mathbf{G}_t, \tag{6}$$

où $\mathbf{G}_t = \left(g_\lambda\left(t\right), \lambda \in m\right)^\top$ et $\overline{\mathbf{\Psi}} = \left(\mathbb{E}\left(a_\lambda a_\mu\right)\right)$ avec $\left(\lambda,\mu\right) \in m \times m$. C'est pourquoi nous cherchons un estimateur $\widehat{\sigma}$ de σ dans la classe des fonctions (6), avec $\mathbf{\Psi} \in \mathbb{R}^{|m|\times|m|}$ une matrice symétrique. Notez que le choix de l'approximation de X (5) dans la base de fonctions, en particulier le choix du sous-ensemble d'indices m, sera cruciale pour les propriétés d'approximation de la fonction de covariance ρ. Cette procédure d'estimation a plusieurs avantages : il sera démontré que le choix approprié de la fonction de perte conduit à la construction d'une matrice symétrique d.n.n. $\widehat{\mathbf{\Psi}}$ et donc l'estimation résultante $\widehat{\sigma}\left(s,t\right) = \mathbf{G}_s^\top \widehat{\mathbf{\Psi}}\mathbf{G}_t$ est une fonction de covariance ; elle peut alors être utilisée dans d'autres procédures qui exigent de travailler avec une fonction de covariance. Nous rappelons également que la grande quantité des approches existantes d'approximation du type (5) (telles que celles basées sur les fonctions de Fourier, ondelettes, les noyaux, ou des fonctions splines) offre une grande flexibilité au modèle (6).

Deuxièmement, nous utilisons la norme de Frobenius pour quantifier le risque de l'estimateur de la matrice de covariance. Nous rappelons que $\mathbf{\Sigma} = \left(\sigma\left(t_j,t_k\right)\right)_{1\leq j,k\leq n}$ est la vraie matrice de covariance alors que $\mathbf{\Gamma} = \left(\rho\left(t_j,t_k\right)\right)_{1\leq j,k\leq n}$ désigne la matrice de covariance de l'approximation $\widetilde{X}$ du processus aux points d'observation. Dès lors, nous obtenons $\mathbf{\Gamma} = \mathbf{G}\overline{\mathbf{\Psi}}\mathbf{G}^\top$. Comparer la fonction de covariance ρ avec la vraie fonction de covariance σ aux points t_j, implique de quantifier la déviation entre $\mathbf{\Gamma}$ et $\mathbf{\Sigma}$. Pour cela, nous considérons la fonction de perte suivante

$$L\left(\mathbf{\Psi}\right) = \mathbb{E}\left\|\mathbf{x}\mathbf{x}^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2,$$

où $\mathbf{x} = (X(t_1), ..., X(t_n))^\top$ et $\|\mathbf{A}\|_F := \sqrt{\operatorname{Tr}(\mathbf{A}^\top \mathbf{A})}$ est la norme de Frobenius définie pour tout matrice $\mathbf{A}$ avec des entrées réelles. Notez que

$$L(\mathbf{\Psi}) = \left\|\mathbf{\Sigma} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + C,$$

où la constante C ne dépend pas de $\mathbf{\Psi}$. À la fonction de perte L correspond le contraste empirique suivant, qui sera le critère que nous allons essayer de minimiser

$$L_N(\mathbf{\Psi}) = \frac{1}{N}\sum_{i=1}^{N} \left\|\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2.$$

Nous rappelons que cette perte est exactement la somme des carrés des résidus correspondant à la régression linéaire du modèle matriciel

$$\mathbf{x}_i\mathbf{x}_i^\top = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, ..., N, \tag{7}$$

avec des matrices d'erreurs $\mathbf{U}_i$ i.i.d. telles que $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$. Cette remarque donne un cadre naturel pour étudier le problème d'estimation de covariance comme un modèle de régression matriciel. Notez également que l'ensemble des matrices $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$ est un sous-espace vectoriel linéaire de $\mathbb{R}^{n\times n}$ lorsque $\mathbf{\Psi}$ varie sur l'espace des matrices symétriques $\mathcal{S}_{|m|}$.

Pour résumer notre approche, nous proposons une procédure d'estimation en deux étapes : dans un premier temps, pour une matrice $\mathbf{G}$ donnée, nous définissons l'estimateur $\widehat{\mathbf{\Psi}}$ de $\mathbf{\Psi}$ par

$$\widehat{\mathbf{\Psi}} = \underset{\mathbf{\Psi}\in\mathcal{S}_{|m|}}{\arg\min}\, L_N(\mathbf{\Psi}),$$

et nous prenons $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ comme l'estimateur de $\mathbf{\Sigma}$. Ainsi, dans un deuxième temps, nous cherchons à sélectionner la meilleure matrice $\mathbf{G} = \mathbf{G}_m$ parmi une collection de candidats $\{\mathbf{G}_m, m \in \mathcal{M}\}$. Pour ce faire, les méthodes et les résultats de la théorie de la sélection du modèle de régression linéaire peuvent être appliqués au contexte actuel. En particulier les résultats de Baraud [7], Comte [27] ou Loubes et Ludena [58] seront utiles dans le traitement de la sélection de modèle dans le cadre (7). Notez que seules des hypothèses sur les moments du processus, et non sur les distributions spécifiques des données, sont impliquées dans notre procédure d'estimation.

Une inégalité oracle pour l'estimation de la covariance

La première partie de cette section décrit les propriétés de l'estimateur des moindres carrés $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ tandis que dans la seconde partie on propose une procédure de sélection pour choisir automatiquement le meilleur estimateur parmi une collection de candidats.

Estimateur des moindres carrés de la covariance

Soit $\mathbf{G}$ la matrice $n \times |m|$ associée à une famille finie de $|m|$ fonctions de base, l'estimateur des moindres carrés de la covariance $\mathbf{\Sigma}$ est défini par

$$\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top = \underset{\mathbf{\Psi}\in\mathcal{S}_{|m|}}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Gamma}\right\|_F^2 : \mathbf{\Gamma} = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\}, \tag{8}$$

où l'estimateur correspondant de la fonction de covariance σ est $\widehat{\sigma}(s,t) = \mathbf{G}_s^\top \widehat{\mathbf{\Psi}} \mathbf{G}_t$.

Proposition 1. *Soit* $\mathbf{Y}_1, ..., \mathbf{Y}_N \in \mathbb{R}^{n\times n}$ *et* $\mathbf{G} \in \mathbb{R}^{n\times |m|}$ *des matrices arbitraires. Alors,*

(a) L'infimum

$$\inf_{\mathbf{\Psi}\in\mathcal{S}_{|m|}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{Y}_i - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|_F^2 \right\}$$

est atteint en

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^- \mathbf{G}^\top \left(\frac{\overline{\mathbf{Y}} + \overline{\mathbf{Y}}^\top}{2} \right) \mathbf{G} \left(\mathbf{G}^\top\mathbf{G}\right)^-,$$

où $\left(\mathbf{G}^\top\mathbf{G}\right)^-$ *est une inverse généralisée de* $\mathbf{G}^\top\mathbf{G}$ *(voir Seber [84] pour une définition générale), et* $\overline{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Y}_i$.

(b) De plus, $\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ *est la même matrice pour toutes les inverses généralisées* $\left(\mathbf{G}^\top\mathbf{G}\right)^-$ *de* $\mathbf{G}^\top\mathbf{G}$*. En particulier, si* $\mathbf{Y}_1, ..., \mathbf{Y}_N \in \mathcal{S}_n$ *(i.e., si ce sont des matrices symétriques) alors tout minimiseur est de la forme*

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^- \mathbf{G}^\top \overline{\mathbf{Y}} \mathbf{G} \left(\mathbf{G}^\top\mathbf{G}\right)^-.$$

Si $\mathbf{Y}_1, ..., \mathbf{Y}_N$ *sont d.n.n. alors la matrice* $\widehat{\mathbf{\Psi}}$ *est d.n.n.*

Théorème 3. *Soit* $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top$*. Alors, l'estimateur de moindres carrés de la covariance définie par (8) est donné par la matrice d.n.n.* $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top = \mathbf{\Pi}\mathbf{S}\mathbf{\Pi}$*, où* $\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^- \mathbf{G}^\top\mathbf{S}\mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^-$ *et* $\mathbf{\Pi} = \mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^-\mathbf{G}^\top$*.*

La preuve de ce théorème est une application directe de la Proposition 1. Ainsi pour une matrice $\mathbf{G}$, l'estimateur des moindres carrés est bien défini et a la structure d'une matrice de covariance. Il reste à étudier la façon de choisir automatiquement l'estimateur lorsqu'est disponible une collection de matrices $\{\mathbf{G}_m : m \in \mathcal{M}\}$ venant de plusieurs choix d'approximations du processus X.

Résultat principal

Nous considérons une collection d'indices $m \in \mathcal{M}$ de taille $|m|$. Soit $\{\mathbf{G}_m : m \in \mathcal{M}\}$ une famille finie de matrices $\mathbf{G}_m \in \mathbb{R}^{n\times|m|}$, et soit $\widehat{\mathbf{\Sigma}}_m = \widehat{\mathbf{\Sigma}}(\mathbf{G}_m)$, $m \in \mathcal{M}$, les estimateurs des moindres carrés de la covariance correspondante. Le problème devient alors de sélectionner le meilleur de ces estimateurs au sens du risque quadratique minimal $\mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2$.

Le théorème principal de cette section donne une majoration non-asymptotique pour le risque d'une stratégie pénalisée pour ce problème. Pour tout $m \in \mathcal{M}$, nous notons $\mathbf{\Pi}_m = \mathbf{G}_m\left(\mathbf{G}_m^\top\mathbf{G}_m\right)^-\mathbf{G}_m^\top$ et $D_m = \mathrm{Tr}\left(\mathbf{\Pi}_m\right)$. Nous supposons que $D_m \geq 1$ pour tout $m \in \mathcal{M}$. L'erreur d'estimation pour un modèle donné $m \in \mathcal{M}$ est donnée par

$$\mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2 = \left\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\right\|_F^2 + \frac{\delta_m^2 D_m}{N}, \tag{9}$$

où $\delta_m^2 = \frac{\mathrm{Tr}((\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m)\mathbf{\Phi})}{D_m}$, $\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$, $\mathbf{A} \otimes \mathbf{B}$ désigne le produit de Kronecker entre les matrices $\mathbf{A}$ et $\mathbf{B}$, $vec\,(\mathbf{A})$ désigne le vecteur obtenu en mettant les colonnes de la matrice $\mathbf{A}$ les unes au dessus des autres, et la matrice $\mathbb{V}(\mathbf{z}) = (Cov\,(Z_i, Z_j))_{1 \le i,j \le n^2}$, où le vecteur $\mathbf{z} = (Z_i)_{1 \le i \le n^2}$.

Étant donné $\theta > 0$, nous définissons l'estimateur pénalisé de la covariance $\widetilde{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_{\widehat{m}}$ par

$$\widehat{m} = \underset{m \in \mathcal{M}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i \mathbf{x}_i^\top - \widehat{\mathbf{\Sigma}}_m \right\|_F^2 + pen\,(m) \right\},$$

où $pen\,(m) = (1+\theta)\,\frac{\delta_m^2 D_m}{N}$. Le théorème suivant est alors vérifié :

Théorème 4. *Soit $q > 0$ tel qu'il existe $p > 2\,(1+q)$ satisfaisant $\mathbb{E}\left\|\mathbf{x}_1\mathbf{x}_1^\top\right\|_F^p < \infty$. Alors, pour des constantes $K\,(\theta) > 1$ et $C'\,(\theta, p, q) > 0$ nous avons*

$$\left(\mathbb{E}\left\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\right\|_F^{2q}\right)^{1/q} \le 2^{\left(q^{-1}-1\right)_+} \left[K\,(\theta) \inf_{m \in \mathcal{M}} \left(\left\|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\right\|_F^2 + \frac{\delta_m^2 D_m}{N}\right) + \frac{\Delta_p}{N}\delta_{\sup}^2\right],$$

où

$$\Delta_p^q = C'\,(\theta, p, q)\,\mathbb{E}\left\|\mathbf{x}_1\mathbf{x}_1^\top\right\|_F^p \left(\sum_{m \in \mathcal{M}} \delta_m^{-p} D_m^{-(p/2-1-q)}\right)$$

et $\delta_{\sup}^2 = \max\{\delta_m^2 : m \in \mathcal{M}\}$. En particulier, pour $q = 1$ nous avons

$$\mathbb{E}\left\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\right\|^2 \le K\,(\theta) \inf_{m \in \mathcal{M}} \mathbb{E}\left(\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2\right) + \frac{\Delta_p}{N}\delta_{\sup}^2. \tag{10}$$

Remarque. *Notez que la pénalité dépend de la quantité δ_m qui est inconnue dans la pratique. En effet, δ_m dépend de $\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$, qui reflète la structure de corrélation des données. Dans la pratique, lorsque N est assez grand, cette quantité peut être estimée en utilisant la version empirique de $\mathbf{\Phi}$ puisque les variables aléatoires $\mathbf{x}_i$, $i = 1, \ldots, N$ observées sont i.i.d. Cet estimateur est donné par*

$$\widehat{\mathbf{\Phi}} = \frac{1}{N}\sum_{i=1}^{N} vec\left(\mathbf{x}_i\mathbf{x}_i^\top\right)\left(vec\left(\mathbf{x}_i\mathbf{x}_i^\top\right)\right)^\top - vec(\mathbf{S})\,(vec(\mathbf{S}))^\top.$$

Par conséquent, il existe une façon pratique de calculer la pénalité et ainsi de construire l'estimateur.

Nous avons par conséquent obtenu dans le Théorème 4 une inégalité oracle, car, en utilisant (9) et (10), on voit immédiatement que $\widetilde{\mathbf{\Sigma}}$ a le même risque quadratique que celui de l'estimateur “oracle” à l'exception d'un terme additif de l'ordre $O\left(\frac{1}{N}\right)$ et un facteur constant. Par conséquent, la procédure de sélection est optimale dans le sens où elle se comporte comme si le vrai modèle était connu.

Estimation de type Group-Lasso de la matrice de covariance dans un cadre de grande dimension

Soit T un sous-ensemble de $\mathbb{R}^d$, avec $d \in \mathbb{N}$. Soit $X = \{X(t),\ t \in T\}$ un processus stochastique à valeurs dans $\mathbb{R}$ de moyenne égale à zéro pour tout $t \in T$, et de fonction de covariance finie $\sigma(s,t) = \mathbb{E}(X(s)X(t))$ pour tous $s,t \in T$. Soit $t_1, \ldots, t_n$ des points fixes dans T (observations déterministes), des copies $X_1, ..., X_N$ indépendantes du processus X, et supposons que nous observons le processus avec du bruit

$$\widetilde{X}_i(t_j) = X_i(t_j) + \mathcal{E}_i(t_j) \ \text{ pour } i = 1, ..., N,\ j = 1, ..., n, \tag{11}$$

où $\mathcal{E}_1, ..., \mathcal{E}_N$ sont des copies indépendantes d'un processus Gaussien du second ordre $\mathcal{E}$ de moyenne nulle et indépendant de X, qui représentent une source de bruit additif dans les mesures. Sur la base des observations bruitées (11), un problème statistique important est la construction d'un estimateur de la matrice de covariance $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ du processus X aux points d'observation, où $\mathbf{X} = (X(t_1), ..., X(t_n))^\top$.

Précédemment, nous avons proposé de construire un estimateur de la matrice de covariance $\mathbf{\Sigma}$ en utilisant N copies indépendantes du processus X, et en projetant le processus X dans un dictionnaire de fonctions de base. La méthode retenue repose sur des techniques de sélection de modèle par la minimisation de contraste empirique dans un modèle de régression matriciel adéquat. Cette nouvelle approche d'estimation de covariance est bien adaptée au cadre de l'estimation de la covariance en basse dimension lorsque le nombre de répétitions du processus N est plus grand que le nombre des points d'observations n. Toutefois, de nombreux domaines d'application sont actuellement aux prises avec le problème de l'estimation d'une matrice de covariance lorsque le nombre d'observations disponibles est faible, comparé au nombre de paramètres à estimer. Les exemples incluent l'imagerie biomédicale, les données génomiques, le traitement du signal dans les neurosciences et bien d'autres. Cette question correspond au problème de l'estimation de la covariance des données en grande dimension. Ce problème est difficile car, dans un cadre de grande dimension (où $n >> N$ ou $n \sim N$), il est bien connu que les matrices de covariance empirique

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i\mathbf{X}_i^\top \in \mathbb{R}^{n\times n},\ \text{où } \mathbf{X}_i = (X_i(t_1), ..., X_i(t_n))^\top, i = 1, \ldots, N$$

et

$$\widetilde{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top \in \mathbb{R}^{n\times n},\ \text{où } \widetilde{\mathbf{X}}_i = \left(\widetilde{X}_i(t_1), ..., \widetilde{X}_i(t_n)\right)^\top, i = 1, \ldots, N$$

ne sont pas des estimateurs convergents de $\mathbf{\Sigma}$. Par exemple, supposons que les $\mathbf{X}_i$ sont des vecteurs aléatoires dans $\mathbb{R}^n$ indépendants et identiquement distribués tirés à partir d'une distribution Gaussienne multivariée. Lorsque $\frac{n}{N} \to c > 0$ quand $n, N \to +\infty$, ni les valeurs propres, ni les vecteurs propres de la matrice de covariance empirique $\mathbf{S}$ ne sont des estimateurs consistants des valeurs propres et vecteurs propres de $\mathbf{\Sigma}$ (voir Johnstone [49]).

Ce sujet a ainsi récemment reçu beaucoup d'attention dans la littérature statistique. Pour assurer la consistance, des méthodes récemment développées pour l'estimation de la covariance en grande dimension imposent des restrictions de sparsité sur la matrice $\mathbf{\Sigma}$. Ces restrictions impliquent que la vraie (mais inconnue) dimension du modèle est beaucoup plus faible que le nombre $\frac{n(n+1)}{2}$ de paramètres d'une matrice de covariance sans contraintes. Sous diverses hypothèses de sparsité, différentes méthodes de régularisation de la matrice de covariance empirique ont été proposées. Des estimateurs basés sur le seuillage des entrées de la matrice de covariance empirique ont été étudiés dans Bickel et Levina [11] et [12]. Le seuillage des composants de la matrice de covariance empirique a également été proposé par El Karoui [38] et la consistance de ces estimateurs est étudiée en utilisant des outils de la théorie des matrices aléatoires. Fan, Fan et Lv [41] imposent des hypothèses de sparsité de la covariance en utilisant un modèle à facteur qui est approprié pour les applications financières. Levina, Rothman et Zhu [57], et Rothman, Bickel, Levina et Zhu [79] proposent des techniques de régularisation avec une pénalité Lasso pour estimer la matrice de covariance ou son inverse. Des pénalisations plus générales ont été étudiées dans Lam et Fan [56].

Une autre approche consiste à imposer de la sparsité sur les vecteurs propres de la matrice de covariance qui mène à une ACP sparse. Zou, Hastie et Tibshirani [93] utilisent une pénalité Lasso pour atteindre une représentation sparse dans l'ACP, d'Aspremont, Bach et El Ghaoui [30] étudient des propriétés de sparsité des composantes principales et de la programmation convexe, tandis que Johnstone et Lu [50] proposent une régularisation de l'ACP en projettant les vecteurs propres empiriques dans une base bien adaptée et ensuite en appliquant une étape de seuillage.

Dans ce chapitre, nous proposons d'estimer $\mathbf{\Sigma}$ dans un cadre de grande dimension en supposant que le processus X a une représentation sparse dans un dictionnaire de fonctions de base. En utilisant un modèle de régression matriciel comme dans la partie précédente, nous proposons une nouvelle méthodologie pour l'estimation de la matrice de covariance en grande dimension, basée sur la régularisation d'un contraste empirique par une pénalisation Group-Lasso. En utilisant une telle pénalité, la méthode sélectionne un ensemble sparse de fonctions de base dans le dictionnaire utilisé pour approcher le processus X. Ceci fournit une approximation de la matrice de covariance $\mathbf{\Sigma}$ dans un espace de plus faible dimension, et donc nous obtenons une nouvelle méthode de réduction de la dimension pour données de grande dimension. Les estimateurs Group-Lasso ont été étudiés dans le modèle linéaire standard et dans le cadre de l'apprentissage à noyaux multiples pour imposer une structure de sparsité par groupe sur les paramètres à récupérer (voir Nardi et Rinaldo [70] et Bach [5]). Toutefois, cette méthodologie n'a pas été utilisée pour l'estimation des matrices de covariance en utilisant une approximation fonctionnelle du processus X.

Cadre statistique

Pour imposer des restrictions de sparsité sur la matrice de covariance $\mathbf{\Sigma}$, notre approche est également basée sur une approximation du processus dans un dictionnaire fini de fonctions de base (pas nécessairement orthogonales) $g_m : T \to \mathbb{R}$ pour $m = 1, ..., M$.

Supposons que

$$X\left(t\right) \approx \sum_{m=1}^{M} a_m g_m\left(t\right), \tag{12}$$

où a_m, $m = 1, ..., M$ sont des variables aléatoires à valeurs réelles, et que pour chaque trajectoire X_i,

$$X_i\left(t_j\right) \approx \sum_{m=1}^{M} a_{i,m} g_m\left(t_j\right). \tag{13}$$

La notation ≈ signifie que le processus X peut être bien approximé dans le dictionnaire. Un sens précis de cette question sera défini plus tard. Alors (13) peut être écrite en notation matricielle comme $\mathbf{X}_i \approx \mathbf{Ga}_i$, $i = 1, ..., N$, où $\mathbf{G}$ est la matrice $n \times M$ dont les entrées sont $\mathbf{G}_{jm} = g_m\left(t_j\right)$ avec $1 \leq j \leq n$ et $1 \leq m \leq M$, et $\mathbf{a}_i$ est le $M \times 1$ vecteur aléatoire de composantes $a_{i,m}$, avec $1 \leq m \leq M$. Étant donné $\mathbf{X} \approx \mathbf{Ga}$ avec $\mathbf{a} = \left(a_m\right)_{1 \leq m \leq M}$ et a_m comme dans (12), alors

$$\mathbf{\Sigma} \approx \mathbb{E}\left(\mathbf{Ga}\left(\mathbf{Ga}\right)^{\top}\right) = \mathbb{E}\left(\mathbf{Gaa}^{\top}\mathbf{G}^{\top}\right) = \mathbf{G}\mathbf{\Psi}^{*}\mathbf{G}^{\top} \text{ avec } \mathbf{\Psi}^{*} = \mathbb{E}\left(\mathbf{aa}^{\top}\right).$$

Nous considérons le modèle de régression matriciel suivant

$$\widetilde{\mathbf{S}} = \mathbf{\Sigma} + \mathbf{U} + \mathbf{W}, \tag{14}$$

où $\mathbf{U} \in \mathbb{R}^{n \times n}$ est une matrice d'erreur centrée donnée par $\mathbf{U} = \mathbf{S} - \mathbf{\Sigma}$ et $\mathbf{W} = \frac{1}{N}\sum\limits_{i=1}^{N} \mathbf{W}_i$, où $\mathbf{W}_i = \mathcal{E}_i \mathcal{E}_i^{\top} \in \mathbb{R}^{n \times n}$ et $\mathcal{E}_i = \left(\mathcal{E}_i\left(t_1\right), ..., \mathcal{E}_i\left(t_n\right)\right)^{\top}$, $i = 1, \ldots, N$.

La taille M du dictionnaire peut être très grande, mais nous faisons l'hypothèse que le processus X a une écriture sparse dans cette base, ce qui signifie que, dans l'approximation (12), un grand nombre de coefficients aléatoires a_m sont proches de zéro. Nous avons obtenu une estimation de la covariance $\mathbf{\Sigma}$ sous la forme $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^{\top}$, telle que $\widehat{\mathbf{\Psi}}$ est une matrice $M \times M$ symétrique avec de nombreuses lignes égales à zéro (et donc, par symétrie, de nombreux colonnes nulles). Notez que si la ligne k-ième de $\widehat{\mathbf{\Psi}}$ est égale à $\mathbf{0} \in \mathbb{R}^{M}$, alors la fonction g_k de l'ensemble des fonctions de base $\left(g_m\right)_{1 \leq m \leq M}$ est supprimée dans la projection sur la base de fonctions associées à $\mathbf{G}$. Pour sélectionner un ensemble sparse de lignes / colonnes de la matrice $\widehat{\mathbf{\Psi}}$ nous utilisons une approche de type Group-Lasso. Pour cela, nous définissons l'estimateur Group-Lasso de la matrice de covariance $\mathbf{\Sigma}$ par

$$\widehat{\mathbf{\Sigma}}_{\lambda} = \mathbf{G}\widehat{\mathbf{\Psi}}_{\lambda}\mathbf{G}^{\top} \in \mathbb{R}^{n \times n}, \tag{15}$$

où $\widehat{\mathbf{\Psi}}_{\lambda}$ est la solution du problème d'optimisation suivant :

$$\widehat{\mathbf{\Psi}}_{\lambda} = \underset{\mathbf{\Psi} \in \mathcal{S}_M}{\arg\min} \left\{ \left\| \widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^{\top} \right\|_F^2 + 2\lambda \sum_{k=1}^{M} \gamma_k \sqrt{\sum_{m=1}^{M} \Psi_{mk}^2} \right\}, \tag{16}$$

où $\mathbf{\Psi} = \left(\Psi_{mk}\right)_{1 \leq m,k \leq M}$ est une $M \times M$ matrice symétrique, λ est un paramètre de régularisation positive et γ_k sont des poids appropriés. Notez que le terme de pénalité impose de donner la préférence à des solutions avec des composantes $\mathbf{\Psi}_k = \mathbf{0}$, où $(\mathbf{\Psi}_k)_{1\leq k\leq M}$ désigne les colonnes de $\mathbf{\Psi}$. Ainsi $\widehat{\mathbf{\Psi}}_\lambda \in \mathbb{R}^{M\times M}$ peut être interprété comme l'estimateur Group-Lasso de $\mathbf{\Sigma}$ dans le modèle de régression matriciel (14).

Avant de présenter les principaux résultats dans la prochaine section, nous rappelons quelques définitions. Pour une matrice $\mathbf{A}$ symétrique avec des entrées réelles de dimension $n\times n$, $\rho_{\min}(\mathbf{A})$ désigne la plus petite valeur propre de $\mathbf{A}$, et $\rho_{\max}(\mathbf{A})$ désigne la plus grande valeur propre de $\mathbf{A}$. Pour $\beta \in \mathbb{R}^q$, $\|\beta\|_{\ell_2}$ désigne la norme euclidienne usuelle de β. Pour une matrice $n\times\ q$ avec des entrées réelles, $\|\mathbf{A}\|_2 = \sup_{\beta\in\mathbb{R}^q,\,\beta\neq 0} \frac{\|\mathbf{A}\beta\|_{\ell_2}}{\|\beta\|_{\ell_2}}$ désigne la norme d'opérateur de $\mathbf{A}$. Rappelons que si $\mathbf{A}$ est une matrice définie non-négative avec $n = q$ alors $\|\mathbf{A}\|_2 = \rho_{\max}(\mathbf{A})$.

Soit $\mathbf{\Psi} \in \mathcal{S}_M$ et β un vecteur de $\mathbb{R}^M$. Pour un sous-ensemble $J \subset \{1,\ldots,M\}$ d'indices de cardinal $|J|$, β_J est le vecteur de $\mathbb{R}^M$ qui a les mêmes coordonnées que β sur J et des coordonnées égales à zéro sur le complémentaire J^c de J. La matrice $n\times|J|$ obtenue en supprimant les colonnes de $\mathbf{G}$ dont les indices ne sont pas en J est représentée par $\mathbf{G}_J$. La sparsité de $\mathbf{\Psi}$ est définie comme le nombre de colonnes non nulles (et donc par le nombre de lignes non nulles), à savoir :

Définition 1. *Pour $\mathbf{\Psi} \in \mathcal{S}_M$, la sparsité de $\mathbf{\Psi}$ est définie par la cardinalité de l'ensemble $\mathcal{M}(\mathbf{\Psi}) = \{k : \mathbf{\Psi}_k \neq \mathbf{0}\}$.*

Puis, nous introduisons les quantités suivantes qui contrôlent les valeurs propres minimales de sous-matrices de petite taille extraites de la matrice $\mathbf{G}^\top\mathbf{G}$ et les corrélations entre les colonnes de $\mathbf{G}$:

Définition 2. *Soit $0 < s \leq M$. Alors,*

$$\rho_{\min}(s) := \inf_{\substack{J \subset \{1,\ldots,M\} \\ |J|\leq s}} \left(\frac{\beta_J^\top \mathbf{G}^\top\mathbf{G}\beta_J}{\|\beta_J\|_{\ell_2}^2}\right) = \inf_{\substack{J \subset \{1,\ldots,M\} \\ |J|\leq s}} \rho_{\min}\left(\mathbf{G}_J^\top\mathbf{G}_J\right).$$

Définition 3. *La cohérence mutuelle $\theta(\mathbf{G})$ des colonnes $\mathbf{G}_k$, $k = 1,\ldots,M$ de $\mathbf{G}$ est définie comme*

$$\theta(\mathbf{G}) := \max\left\{\left|\mathbf{G}_{k'}^\top\mathbf{G}_k\right|,\ k\neq k',\ 1\leq k,k'\leq M\right\}$$

et

$$\mathbf{G}_{\max}^2 := \max\left\{\|\mathbf{G}_k\|_{\ell_2}^2,\ 1\leq k\leq M\right\}.$$

Pour obtenir les inégalités oracle montrant la consistance de l'estimateur Group-Lasso $\widehat{\mathbf{\Psi}}_\lambda$ les corrélations entre les colonnes de $\mathbf{G}$ (mesurés par $\theta(\mathbf{G})$) ne doivent pas être trop grandes par rapport aux valeurs propres minimales des petites matrices extraites de $\mathbf{G}^\top\mathbf{G}$, ce qui est formulé dans l'hypothèse suivante :

Hypothèse 4. *Soit $c_0 > 0$ une constante et soit $0 < s \leq M$. Alors*

$$\theta(\mathbf{G}) < \frac{\rho_{\min}(s)^2}{c_0\rho_{\max}(\mathbf{G}^\top\mathbf{G})s}.$$

L'Hypothèse 4 est inspirée par des résultats récents de Bickel, Ritov et Tsybakov [13] sur la consistance des estimateurs Lasso dans le modèle de régression nonparamétrique standard en utilisant un dictionnaire de fonctions de base. Dans Bickel et al. [13], une condition générale appelée *hypothèse aux valeurs propres restreintes* est introduite pour contrôler les valeurs propres minimales de la matrice de Gram associée au dictionnaire sur des ensembles de vecteurs sparses.

Maintenant nous précisons la loi du processus stochastique X. Pour cela, rappelons que pour une variable aléatoire réelle Z, la norme Orlicz ψ_α de Z est

$$\|Z\|_{\psi_\alpha} := \inf\left\{C > 0 : \mathbb{E}\exp\left(\frac{|Z|^\alpha}{C^\alpha}\right) \leq 2\right\}.$$

Ces normes sont utiles pour caractériser le comportement de queue des variables aléatoires. En effet, si $\|Z\|_{\psi_\alpha} < +\infty$ alors cela équivaut à l'affirmation selon laquelle il existe deux constantes $K_1, K_2 > 0$ telles que pour tout $x > 0$ (voir par exemple Mendelson et Pajor [69] pour plus de détails sur les normes Orlicz de variables aléatoires)

$$\mathbb{P}\left(|Z| \geq x\right) \leq K_1 \exp\left(-\frac{x^\alpha}{K_2^\alpha}\right).$$

Par conséquent, si $\|Z\|_{\psi_2} < +\infty$ alors on dit que Z a un comportement sous-gaussien et si $\|Z\|_{\psi_1} < +\infty$ alors on dit que Z a un comportement sous-exponentielle.

Dans les sections suivantes, les inégalités oracles pour l'estimateur Group-Lasso seront obtenues sous l'hypothèse suivante sur X :

Hypothèse 5. *Le vecteur aléatoire* $X = \left(X\left(t_1\right), ..., X\left(t_n\right)\right)^\top \in \mathbb{R}^n$ *est tel que :*

(A1) *Il existe* $\rho\left(\mathbf{\Sigma}\right) > 0$ *telle que, pour tous les vecteurs* $\beta \in \mathbb{R}^n$ *avec* $\|\beta\|_{\ell_2} = 1$*, alors* $\left(\mathbb{E}|\mathbf{X}^\top\beta|^4\right)^{1/4} < \rho\left(\mathbf{\Sigma}\right)$.

(A2) *Soit* $Z = \|\mathbf{X}\|_{\ell_2}$. *Il existe* $\alpha \geq 1$ *tel que* $\|Z\|_{\psi_\alpha} < +\infty$.

Dans la section suivante nous donnons des résultats de la consistance de l'estimateur Group-Lasso que nous avons proposé en utilisant des inégalités oracles.

Consistance de l'estimateur Group-Lasso

La consistance de l'estimateur Group-Lasso est d'abord étudiée en utilisant la norme de Frobenius normalisée définie par $\frac{1}{n}\left\|\mathbf{A}\right\|_F^2$ pour une matrice $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Une inégalité oracle pour la norme de Frobenius

Le théorème suivant fournit une inégalité oracle pour l'estimateur Group-Lasso

$$\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}_\lambda\mathbf{G}^\top.$$

Théorème 5. *Supposons que X satisfait l'Hypothèse 5. Soit $\epsilon > 0$ et $1 \leq s \leq \min(n, M)$. Supposons que l'Hypothèse 4 est satisfaite avec $c_0 = 3 + 4/\epsilon$. Considérons l'estimateur Group-Lasso $\widehat{\boldsymbol{\Sigma}}_\lambda$ définie par (15) avec*

$$\gamma_k = 2\|\mathbf{G}_k\|_{\ell_2}\sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)} \text{ et } \lambda = \|\Sigma_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^2$$

pour une constante $\delta > 1$, où $\Sigma_{noise} = \mathbb{E}(\mathbf{W}_1)$. Alors, avec une probabilité d'au moins $1 - M^{1-\delta}$ on a que

$$\frac{1}{n}\left\|\widehat{\boldsymbol{\Sigma}}_\lambda - \boldsymbol{\Sigma}\right\|_F^2 \leq (1+\epsilon) \inf_{\substack{\boldsymbol{\Psi} \in \mathcal{S}_M \\ \mathcal{M}(\boldsymbol{\Psi}) \leq s}} \left(\frac{4}{n}\left\|\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^\top - \boldsymbol{\Sigma}\right\|_F^2 + \frac{8}{n}\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2 + C(\epsilon)\frac{\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G})}{\kappa_{s,c_0}^2}\lambda^2 \frac{\mathcal{M}(\boldsymbol{\Psi})}{n}\right), \tag{17}$$

où

$$\kappa_{s,c_0}^2 = \rho_{\min}(s)^2 - c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})s$$

et $C(\epsilon) = 8\frac{\epsilon}{1+\epsilon}(1 + 2/\epsilon)^2$.

Le premier terme $\frac{1}{n}\left\|\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^\top - \boldsymbol{\Sigma}\right\|_F^2$ de l'inégalité (17) est le biais de l'estimateur $\widehat{\boldsymbol{\Sigma}}_\lambda$. Il reflète la qualité de l'approximation de $\boldsymbol{\Sigma}$ par l'ensemble des matrices de la forme $\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^\top$ avec $\boldsymbol{\Psi} \in \mathcal{S}_M$ et $\mathcal{M}(\boldsymbol{\Psi}) \leq s$. À titre d'exemple, supposons que $X = X^0$, où le processus X^0 a une représentation sparse dans la base de fonctions $(g_m)_{1 \leq m \leq M}$ donnée par

$$X^0(t) = \sum_{m \in J^*} a_m g_m(t), \ t \in T, \tag{18}$$

où $J^* \subset \{1, \ldots, M\}$ est un sous-ensemble d'indices de cardinalité $|J^*| = s^* \leq s$ et a_m, $m \in J^*$ sont des coefficients aléatoires. Alors, puisque $s^* \leq s$ le terme de biais dans (17) est égal à zéro.

Le second terme $\frac{1}{n}\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2$ dans (17) est un terme de variance puisque $\mathbf{S}$ estime sans biais $\boldsymbol{\Sigma}$. En utilisant $\frac{1}{n}\|\mathbf{A}\|_F^2 \leq \|\mathbf{A}\|_2^2$, qui est vrai pour toute matrice $\mathbf{A} \in \mathbb{R}^{n \times n}$ on obtient que $\frac{1}{n}\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2 \leq \|\mathbf{S} - \boldsymbol{\Sigma}\|_2^2$. Ainsi, si X a une représentation sparse (i.e. quand $X = X_0$) alors le terme de variance $\frac{1}{n}\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2$ est contrôlé par $\frac{s^*}{N} \leq \frac{s}{N}$, et non par $\frac{n}{N}$ sans hypothèse sur la structure de $\boldsymbol{\Sigma}$.

Le troisième terme dans (17) est aussi un terme de variance dû au bruit dans les mesures (11). S'il existe une constante $c > 0$ indépendante de n et N telle que $\frac{n}{N} \leq c$, alors ce troisième terme de variance est essentiellement contrôlé par le ratio $\frac{\mathcal{M}(\boldsymbol{\Psi})}{n} \leq \frac{s}{n}$. Par conséquent, si $\mathcal{M}(\boldsymbol{\Psi}) \leq s$ avec sparsité s beaucoup plus petite que n, alors la variance de l'estimateur Group-Lasso $\widehat{\boldsymbol{\Sigma}}_\lambda$ est plus petite que la variance de $\widetilde{\mathbf{S}}$. Cela montre quelques-unes des améliorations apportées par la régularisation (16) sur la matrice de covariance empirique $\widetilde{\mathbf{S}}$ avec une pénalité Group-Lasso.

Une inégalité oracle pour la norme d'opérateur

La norme de Frobenius normalisée $\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_{\lambda}-\mathbf{\Sigma}\right\|_{F}^{2}$ (la moyenne des valeurs propres) peut être considérée comme un approximation raisonnable de la norme d'opérateur $\left\|\widehat{\mathbf{\Sigma}}_{\lambda}-\mathbf{\Sigma}\right\|_{2}^{2}$. Il est donc attendu que les résultats du Théorème 5 impliquent que l'estimateur Group-Lasso $\widehat{\mathbf{\Sigma}}_{\lambda}$ est un bon estimateur de $\mathbf{\Sigma}$ dans la norme d'opérateur. Pour le voir, nous considérons le cas où X consiste en des observations bruitées du processus X^0 (18), ce qui signifie que

$$\widetilde{X}(t_j) = X^0(t_j) + \mathcal{E}\left(t_j\right),\ j = 1, \ldots, n, \tag{19}$$

où $\mathcal{E}$ est un processus Gaussien de second ordre de moyenne nulle et indépendant de X^0. Dans ce cas, on a que $\mathbf{\Sigma} = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top$, où $\mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right)$ et $\mathbf{a}$ est le vecteur aléatoire de $\mathbb{R}^M$ avec $\mathbf{a}_m = a_m$ pour $m \in J^*$ et $\mathbf{a}_m = 0$ pour $m \notin J^*$. Par conséquent, en utilisant le Théorème 5 avec $s = |J^*| = s^*$, *puisque* $\mathbf{\Psi}^* \in \left\{\mathbf{\Psi} \in \mathcal{S}_M : M\left(\mathbf{\Psi}\right) \leq s^*\right\}$, on peut en déduire le corollaire suivant :

Corollaire 1. *Supposons que les observations sont des variables aléatoires i.i.d. du modèle (19) et que les conditions du Théorème 5 sont satisfaites avec* $1 \leq s = s^* \leq \min(n, M)$. *Alors, avec une probabilité d'au moins* $1 - M^{1-\delta}$ *on a que*

$$\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_{\lambda}-\mathbf{\Sigma}\right\|_{F}^{2} \leq C_0\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right), \tag{20}$$

où

$$C_0\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right) = (1+\epsilon)\left(\frac{8}{n}\left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top\right\|_F^2 + C(\epsilon)\frac{\mathbf{G}^2_{\max}\rho_{\max}(\mathbf{G}^\top\mathbf{G})}{\kappa^2_{s^*,c_0}}\lambda^2\frac{s^*}{n}\right).$$

Pour simplifier les notations, écrivons $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_{\lambda}$, avec $\widehat{\mathbf{\Psi}}_{\lambda}$ donnée par (16). Nous définissons $\widehat{J}_{\lambda} \subset \{1, \ldots, M\}$ par

$$\widehat{J}_{\lambda} \equiv \widehat{J} := \left\{k : \frac{\delta_k}{\sqrt{n}}\left\|\widehat{\mathbf{\Psi}}_k\right\|_{\ell_2} > C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)\right\},\ \text{avec } \delta_k = \frac{\left\|\mathbf{G}_k\right\|_{\ell_2}}{\mathbf{G}_{\max}} \text{ et} \tag{21}$$

$$C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right) = \frac{4\left(1+\epsilon\right)\sqrt{s^*}}{\epsilon\kappa_{s^*,c_0}}\sqrt{C_0\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)}. \tag{22}$$

L'ensemble des indices $\widehat{J}$ est une estimation de l'ensemble des fonctions actives dans la base J^*. Notez que pour estimer J^* nous n'avons pas simplement pris $\widehat{J} = \left\{k : \left\|\widehat{\mathbf{\Psi}}_k\right\|_{\ell_2} \neq 0\right\}$, mais nous appliquons plutôt une étape de seuillage pour se débarrasser des colonnes de $\widehat{\mathbf{\Psi}}$ dont la norme ℓ_2 est trop petite. Une étape de seuillage similaire est proposée dans Lounici [61] et Lounici, Pontil, Tsybakov et van de Geer [62] dans le modèle linéaire standard pour sélectionner un ensemble sparse de variables actives lors de l'utilisation de régularisation avec une pénalité Lasso. Le théorème devient :

Théorème 6. *Sous les hypothèses du Corollaire 1, pour toute solution du problème (16), nous avons que, avec une probabilité d'au moins* $1 - M^{1-\delta}$,

$$\max_{1\le k\le M} \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}^*_k \right\|_{\ell_2} \le C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right). \tag{23}$$

De plus, si

$$\min_{k\in J^*} \frac{\delta_k}{\sqrt{n}} \left\| \mathbf{\Psi}^*_k \right\|_{\ell_2} > 2C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right) \tag{24}$$

alors avec la même probabilité l'ensemble d'indices $\widehat{J}$*, définie par (21), estime correctement le vrai ensemble des fonctions de base actives* J^**, c'est à dire,* $\widehat{J} = J^*$ *avec une probabilité d'au moins* $1 - M^{1-\delta}$.

Les résultats du Théorème 6 indiquent que si la norme ℓ_2 des colonnes de $\mathbf{\Psi}^*_k$ pour $k \in J^*$ est suffisamment grande par rapport à l'indice de sparsité s^*, alors $\widehat{J}$ est une estimation consistante de l'ensemble des variables actives. Ceci suggère de prendre comme estimateur final de $\mathbf{\Sigma}$ la matrice $\widehat{\mathbf{\Sigma}}_{\widehat{J}} = \mathbf{G}_{\widehat{J}} \widehat{\mathbf{\Psi}}_{\widehat{J}} \mathbf{G}_{\widehat{J}}$, où $\mathbf{G}_{\widehat{J}}$ dénote la matrice $n \times |\widehat{J}|$ obtenue en supprimant les colonnes de $\mathbf{G}$ dont les indices ne sont pas dans $\widehat{J}$, et

$$\widehat{\mathbf{\Psi}}_{\widehat{J}} = \operatorname*{argmin}_{\mathbf{\Psi}\in\mathcal{S}_{|\widehat{J}|}} \left\{ \left\| \widetilde{\mathbf{S}} - \mathbf{G}_{\widehat{J}} \mathbf{\Psi} \mathbf{G}_{\widehat{J}}^\top \right\|_F^2 \right\},$$

où $\mathcal{S}_{|\widehat{J}|}$ désigne l'ensemble des matrices $|\widehat{J}| \times |\widehat{J}|$ symétriques. Notez que si $\mathbf{G}_{\widehat{J}}^\top \mathbf{G}_{\widehat{J}}$ est inversible, alors

$$\widehat{\mathbf{\Psi}}_{\widehat{J}} = \left(\mathbf{G}_{\widehat{J}}^\top \mathbf{G}_{\widehat{J}}\right)^{-1} \mathbf{G}_{\widehat{J}}^\top \widetilde{\mathbf{S}} \mathbf{G}_{\widehat{J}} \left(\mathbf{G}_{\widehat{J}}^\top \mathbf{G}_{\widehat{J}}\right)^{-1}.$$

Rappelons que si les observations sont des variables aléatoires i.i.d. à partir du modèle (19), alors $\mathbf{\Sigma} = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top$, où $\mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right)$, et $\mathbf{a}$ est le vecteur aléatoire de $\mathbb{R}^M$ avec $\mathbf{a}_m = a_m$ pour $m \in J^*$ et $\mathbf{a}_m = 0$ pour $m \notin J^*$. Alors, nous définissons le vecteur aléatoire $\mathbf{a}_{J^*} \in \mathbb{R}^{J^*}$ dont les coordonnées sont les coefficients aléatoires a_m pour $m \in J^*$. Soit $\mathbf{\Psi}_{J^*} = \mathbb{E}\left(\mathbf{a}_{J^*}\mathbf{a}_{J^*}^\top\right)$ et dénotons par $\mathbf{G}_{J^*}$ la matrice $n \times |J^*|$ obtenue en supprimant les colonnes de $\mathbf{G}$ dont les indices ne sont pas dans J^*. Notez que $\mathbf{\Sigma} = \mathbf{G}_{J^*}\mathbf{\Psi}_{J^*}\mathbf{G}_{J^*}^\top$.

En supposant que $\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}$ est inversible, nous définissons la matrice

$$\mathbf{\Sigma}_{J^*} = \mathbf{\Sigma} + \mathbf{G}_{J^*}(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1} \mathbf{G}_{J^*}^\top \mathbf{\Sigma}_{noise} \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top. \tag{25}$$

Alors, le théorème suivant donne un contrôle de la déviation entre $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ et $\mathbf{\Sigma}_{J^*}$ dans la norme d'opérateur.

Théorème 7. *Supposons que les observations sont des variables aléatoires i.i.d. du modèle (19) et que les conditions du Théorème 5 sont satisfaites avec* $1 \le s = s^* \le \min(n, M)$*. Supposons que* $\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}$ *est une matrice inversible, et que*

$$\min_{k\in J^*} \frac{\delta_k}{\sqrt{n}} \left\| \mathbf{\Psi}^*_k \right\|_{\ell_2} > 2C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right),$$

où $C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)$ *est la constante définie dans (22).*

Soit $\mathbf{Y} = \left(\mathbf{G}_{J^*}^{\top}\mathbf{G}_{J^*}\right)^{-1}\mathbf{G}_{J^*}^{\top}\widetilde{\mathbf{X}}$ *et* $\widetilde{Z} = \|\mathbf{Y}\|_{\ell_2}$. *Soit* $\rho^4\left(\mathbf{\Sigma}_{noise}\right) = \sup_{\beta\in\mathbb{R}^n, \|\beta\|_{\ell_2}=1} \mathbb{E}|\mathcal{E}^{\top}\beta|^4$, où $\mathcal{E} = \left(\mathcal{E}\left(t_1\right), ..., \mathcal{E}\left(t_n\right)\right)^{\top}$.

Alors, avec une probabilité d'au moins $1 - M^{1-\delta} - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$, *avec* $\delta > 1$ *et* $\delta_\star > \delta_*$ *on a que*

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^{\top}\mathbf{G}_{J^*}\right)\widetilde{\tau}_{N,s^*}\delta_\star\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}, \tag{26}$$

où $\widetilde{\tau}_{N,s^*} = \max(\widetilde{A}_{N,s^*}^2, \widetilde{B}_{N,s^*})$, *avec*

$$\widetilde{A}_{N,s^*} = \|\widetilde{Z}\|_{\psi_\alpha}\frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$$

et

$$\widetilde{B}_{N,s^*} = \frac{\widetilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})\rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^{\top}\mathbf{G}_{J^*}\right)}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^{\top}\mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2}\widetilde{A}_{N,s^*},$$

où $d^* = \min(N, s^*)$ *et* $\widetilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) = 8^{1/4}\left(\rho^4\left(\mathbf{\Sigma}\right) + \rho^4\left(\mathbf{\Sigma}_{noise}\right)\right)^{1/4}$.

Notons que le théorème ci-dessus donne une déviation dans la norme d'opérateur de la matrice $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ et la matrice $\mathbf{\Sigma}_{J^*}$ définie dans (25), qui n'est pas égal à la vraie covariance $\mathbf{\Sigma}$ de X aux points d'observations. En effet, même si nous connaissons le vrai ensemble J^*, le bruit additif de mesure dans le modèle (11) complique l'estimation de $\mathbf{\Sigma}$ dans la norme d'opérateur. Cependant, bien que $\mathbf{\Sigma}_{J^*} \neq \mathbf{\Sigma}$, les deux matrices peuvent avoir les mêmes vecteurs propres si la structure du terme de la matrice du bruit additif dans (25) n'est pas trop complexe. À titre d'exemple, prenons le cas d'un bruit blanc additif, pour lequel $\mathbf{\Sigma}_{noise} = \sigma^2\mathbf{I}_n$, où σ est le niveau de bruit et $\mathbf{I}_n$ est la matrice identité. Sous une telle hypothèse, si l'on suppose en outre pour simplifier que $(\mathbf{G}_{J^*}^{\top}\mathbf{G}_{J^*})^{-1} = \mathbf{I}_{s^*}$, alors $\mathbf{\Sigma}_{J^*} = \mathbf{\Sigma} + \sigma^2\mathbf{G}_{J^*}(\mathbf{G}_{J^*}^{\top}\mathbf{G}_{J^*})^{-1}\mathbf{G}_{J^*}^{\top} = \mathbf{\Sigma} + \sigma^2\mathbf{I}_n$ et clairement $\mathbf{\Sigma}_{J^*}$ et $\mathbf{\Sigma}$ ont les mêmes vecteurs propres. Par conséquent, les vecteurs propres de $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ peuvent être utilisés comme estimateurs des vecteurs propres de $\mathbf{\Sigma}$, et $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ est approprié pour les applications des ACP sparse.

Conclusions générales

Dans ce travail nous avons étudié le problème de l'estimation de la covariance. Nous avons proposé différentes méthodes nonparamétriques pour l'estimation de la fonction de covariance et de la matrice de covariance d'un processus stochastique dans des conditions différentes sur le processus. Nous avons étudié les proprietés théoriques des éstimateurs et leurs comportements dans la pratique.

Chapter 1

General Presentation

The main objective of this thesis is to develop nonparametric methods for covariance estimation of a stochastic process. Under different conditions on the process, nonparametric estimators of the covariance function will be introduced, with the property of being non-negative definite functions. Furthermore, a method for the estimation of the covariance matrix of a stochastic process in a high-dimensional setting will be proposed.

1.1 Motivation

Covariance estimation is a fundamental problem in inference for stochastic processes with many applications, ranging from geology, meteorology, financial time series, epidemiology, etc. It is required when an accurate prediction is needed, from a sequence of observations. For example, assume that we are given observations $X\left(t_1\right), ..., X\left(t_n\right)$ of a real-valued stochastic process $\{X\left(t\right) : t \in T\}$, where $T \subset \mathbb{R}^d$ with $d \in \mathbb{N}$, and we wish to map the surface X within the region T. The sample points $t_1, ..., t_n$ are usually points at which the data are available. For instance, networks of rain gauges are set up where observers are available, and data on oil and mineral fields are available where drilling occurred (at spots thought to be fruitful) and from those parts of the field that are being exploited. Such a surface can have independent uses. In mining, a map of the mineral grade (for example, percentage of copper) will help plan the mining operation as well as give information on which parcels will have a high enough average grade to make processing economic. It may also be helpful to have a smoothed map to indicate the broad features of the data as well as an interpolated map for prediction.

Linear inference for one position $t_0 \in T$ is given by the pair $\left(\widehat{X}(t_0), \widehat{\nu}(t_0)\right)$, where the predicted value $\widehat{X}(t_0)$ of $X(t_0)$ is defined as a weighted linear combination of the available observations and $\widehat{\nu}(t_0)$ is the prediction variability associated to $\widehat{X}(t_0)$. The linear method that provides the optimal predictor $\widehat{X}(t_0)$ is called kriging, where the optimality is considered in the sense of minimal prediction variance within the class of unbiased predictors. Kriging is the most popular method used for prediction and bears the name of its inventor, the south african mining engineer D.G. Krige. The predicted value $\widehat{X}(t_0)$ is defined by

$$\widehat{X}(t_0) = \sum_{i=1}^{n} \varpi_i X\left(t_i\right),$$

with $\varpi_1, ..., \varpi_n \in \mathbb{R}$. The unbiasedness condition implies that $\sum\limits_{i=1}^{n} \varpi_i = 1$, and its variance is given by

$$\mathbb{V}\left(\widehat{X}(t_0)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \varpi_i \varpi_j Cov\left(X\left(t_i\right), X\left(t_j\right)\right).$$

Therefore, to ensure that the variance $\mathbb{V}\left(\widehat{X}(t_0)\right)$ is non-negative, the covariance function $\sigma\left(t,s\right) := Cov\left(X\left(t\right), X\left(s\right)\right)$ must be non-negative definite, that is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \sigma\left(t_i, t_j\right) \geq 0 \tag{1.1}$$

for all $n \in \mathbb{N}^*$, $\alpha_1, ..., \alpha_n \in \mathbb{R}$ and $t_1, ..., t_n \in \mathbb{R}^d$, (see Ripley [78] and references therein). The computation of the weights ϖ_i and the prediction variance $\widehat{\nu}(t_0)$ depend on the covariance function, which is unknown in practice and has to be estimated from the available observations. This motivates the search of estimators of the covariance function of a stochastic process satisfying the non-negative definiteness property carried out in this work.

1.1.1 Some preliminary definitions

A stochastic process or random field is a collection of random variables $X = \{X\left(t\right) : t \in T\}$, where $T \subset \mathbb{R}^d$ with $d \in \mathbb{N}$. The Daniell-Kolmogorov extension theorem states that to specify a stochastic process all we have to do is to give the joint distributions of any finite subset $\{X\left(t_1\right), ..., X\left(t_n\right)\}$ in a consistent way, requiring

$$\mathbb{P}\left(X\left(t_i\right) \in A_i,\ i = 1, ..., m,\ X\left(s\right) \in \mathbb{R} \text{ for } s = m+1, ..., n\right) = \mathbb{P}\left(X\left(t_i\right) \in A_i,\ i = 1, ..., m\right).$$

Such a specification is called the distribution of the process. A random field X is called Gaussian if, for all $n \in \mathbb{N}$ and for all $t_1, ..., t_n$, the vector $\left(X\left(t_1\right), ..., X\left(t_n\right)\right)^\top$ has a Normal distribution. Suppose that the variance of $X\left(t\right)$ exists for all $t \in T$. Let $\mu\left(t\right) = \mathbb{E}\left(X\left(t\right)\right)$ be the mean of the process X at position t.

We say that the stochastic process X defined on $T = \mathbb{R}^d$ is strictly stationary if its distribution is unchanged when the origin of the index set is translated, that is, if the distribution of the vector $\left(X\left(t_1\right), ..., X\left(t_n\right)\right)^\top$ is the same that the distribution of $\left(X\left(t_1+h\right), ..., X\left(t_n+h\right)\right)^\top$ for all $h \in \mathbb{R}^d$ and for all $t_1, ..., t_n$ in T such that $t_1+h, ..., t_n+h$ belong to T. We also say that X is stationary of order m if

$$\mathbb{E}\left(\left[X\left(t_1\right)\right]^{m_1} \cdots \left[X\left(t_n\right)\right]^{m_n}\right) = \mathbb{E}\left(\left[X\left(t_1+h\right)\right]^{m_1} \cdots \left[X\left(t_n+h\right)\right]^{m_n}\right)$$

for all $h \in T$ and for all positive integers $m_1, ..., m_n$ such that $\sum\limits_{i=1}^{n} m_i \leq m$. If the process X has moments of order 2, taking $m = 2$ we get

$$\mathbb{E}\left(X\left(t\right)\right) = \mu, \text{ for all } t \in \mathbb{R}^d$$

and

$$Cov\left(X\left(t\right),X\left(s\right)\right)=\sigma\left(t-s\right),\ \text{for all}\ \left(t,s\right)\in\mathbb{R}^{2d}$$

for some function σ defined on T. In this case the process X is called stationary of second order and the function σ is called covariogram or stationary covariance function. The random field X is called isotropic if it is stationary of second order and the covariance between two observations $X\left(t\right)$ and $X\left(s\right)$ is a function of the distance between the points t and s, that is

$$Cov\left(X\left(t\right),X\left(s\right)\right)=\sigma_0\left(\left\|t-s\right\|_{\ell_2}\right),\ \text{for all}\ \left(t,s\right)\in\mathbb{R}^{2d},$$

where σ_0 is a function defined on $\mathbb{R}_+$ and $\|.\|_{\ell_2}$ is the Euclidean norm in $\mathbb{R}^d$.

1.1.2 Characterization and spectral representation of stationary covariance functions

A necessary and sufficient condition for a continuous function σ to be a stationary covariance function is that σ is non-negative definite (see 1.1). In the case of a stationary process X with continuous covariance function σ an alternative to this characterization is given by Bochner's theorem [20], which states that a continuous function σ on $\mathbb{R}^d$ is non-negative definite if and only if it is the Fourier transform of a bounded non-negative measure called the spectral measure. When a spectral measure has a density, this density is called the spectral density and is defined by

$$f\left(\omega\right)=\frac{1}{\left(2\pi\right)^d}\int_{\mathbb{R}^d}\sigma\left(h\right)\exp\left(-i\omega^\top h\right)dh. \tag{1.2}$$

Then

$$\sigma\left(h\right)=\int_{\mathbb{R}^d}f\left(\omega\right)\exp\left(i\omega^\top h\right)d\omega. \tag{1.3}$$

Therefore, the necessary and sufficient condition for a continuous function σ to be a stationary covariance function (or equivalently a non-negative definite function) is that $f\left(\omega\right)\geq 0$ for all ω. For processes on a lattice, in (1.2) the integral is replaced by a sum, and only frequencies for which each component is in the range $[-\pi,\pi]$ are considered, so the integration in (1.3) is restricted to $[-\pi,\pi]^d$. Any non-negative function that gives a finite value of $\sigma\left(0\right)$ in (1.3) is a spectral density.

1.1.3 Classical covariance estimation methods

Our main references for this section are the book of Cressie [28] and the introduction of the thesis of Elogne [39].

Let $X(t_1),X(t_2),...,X(t_n)$ be real observations of the process X at points $t_1,t_2,...,t_n$ in $\mathbb{R}^d$. Under stationarity, the classical approach to estimate the covariance function is the unbiased estimator proposed by Matheron [66], which is obtained by the method of moments. It is defined by

$$\widehat{\sigma}\left(t\right)=\frac{1}{\left|\mathcal{N}\left(t\right)\right|}\sum_{\left(t_i,t_j\right)\in\mathcal{N}\left(t\right)}\left(X(t_i)-\widehat{\mu}_n\right)\left(X(t_j)-\widehat{\mu}_n\right) \tag{1.4}$$

if the points $t_1, ..., t_n$ are on a regular grid, where $|\mathcal{N}(t)|$ is the cardinality of the set $\mathcal{N}(t) = \{(t_i, t_j) : t_i - t_j = t,\ 1 \leq i, j \leq n\}$ and $\widehat{\mu}_n$ is the empirical estimator of the mean of the process μ, defined by

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X(t_i).$$

In the case where the points $t_1, ..., t_n$ are not on a regular grid, the estimator $\widehat{\sigma}(t)$ is computed replacing $\mathcal{N}(t)$ by $\mathcal{N}'(t) = \{(t_i, t_j) : t_i - t_j \in \mathcal{T}(t),\ 1 \leq i, j \leq n\}$, where $\mathcal{T}(t)$ is some tolerance region around t.

These estimators do not satisfy the non-negative definiteness property and therefore generate negative prediction variances. There are several estimation methods based on $\widehat{\sigma}$ that attempt to obtain non-negative definite estimators of the covariance function σ (see the works of Christakos [26] or Shapiro and Botha [85] for instance). This is the case of parametric methods (Cressie [28]), which suppose that the covariance function σ belongs to a parametric family of non-negative definite functions $\left\{\sigma(\theta; \cdot) : \theta \in \Theta \subset \mathbb{R}^d\right\}$. Then, the covariance function estimator is defined as $\sigma\left(\widehat{\theta}; \cdot\right)$, where $\widehat{\theta}$ is some estimator of the parameter θ. For example, the least squares estimator of θ is given by

$$\widehat{\theta} = \underset{\theta \in \Theta = \mathbb{R}^d}{\arg\min} \left\{ (\Delta - \widetilde{\sigma}(\theta))^\top (\Delta - \widetilde{\sigma}(\theta)) \right\},$$

where $\Delta = (\widehat{\sigma}(t_1), ..., \widehat{\sigma}(t_n))^\top$ with $\widehat{\sigma}(t_i)$ given by (1.4) for all $i = 1, ..., n$, and $\widetilde{\sigma}(\theta) = (\sigma(\theta; t_1), ..., \sigma(\theta; t_n))^\top$.

The most used parametric models for covariance functions of isotropic process are listed below. The term valid will be use when the non-negative definiteness property is satisfied.

Model	**Form of** $\sigma(\theta; \cdot)$ **with** $\theta > 0$	**Validity**
Spherical	$\sigma(\theta; t) = \begin{cases} c\left(1 - 1.5\frac{t}{\theta} + 0.5\frac{t^3}{\theta^3}\right), & \text{if } t \leq \theta \\ 0, & \text{if } t \geq \theta \end{cases}$	$\mathbb{R}^d$ for $d \leq 3$
Exponential	$\sigma(\theta; t) = c \exp\left(-\frac{t}{\theta}\right)$	$\mathbb{R}^d$
Gaussian	$\sigma(\theta; t) = c \exp\left(-\frac{t^2}{\theta^2}\right)$	$\mathbb{R}^d$
Cauchy	$\sigma(\theta; t) = c\left(1 + \left(\frac{t}{\theta}\right)^\alpha\right)^{-\frac{\beta}{\alpha}}$, $0 < \alpha \leq 2$, $\beta > 0$	$\mathbb{R}^d$
Bessel	$\sigma(\theta; t) = c\Gamma\left(\frac{\upsilon}{2}\right)\left(\frac{2\theta}{t}\right)^{\frac{(\upsilon-2)}{2}} J_{\frac{(\upsilon-2)}{2}}\left(\frac{t}{\theta}\right)$, $\upsilon \geq 1$	$\mathbb{R}^d$ for $\upsilon \geq d$
Whittle-Matérn	$\sigma(\theta; t) = c\frac{2^{\upsilon-1}}{\Gamma(\upsilon)}\left(\frac{t}{\theta}\right)^\upsilon K_\upsilon\left(\frac{t}{\theta}\right)$, $\upsilon > 0$	$\mathbb{R}^d$

Table 1. Parametric covariance functions. In all cases $c > 0$, Γ is the Gamma function, J_υ is the Bessel function of first kind of order υ and K_υ is the modified Bessel function of second kind of order υ.

In practice, the true covariance model is unknown, therefore the statistician must select a model among the existing ones and compare it with others according to some criterion to ensure the selection of the best model. Note that the correlation functions in Table 1 can be separated into two groups, one containing functions that have a parabolic behavior near the origin (for instance the Gaussian model for instance), and the other containing functions with a linear behaviour close to the origin (Exponential and Spherical models). Thus, the choice of the correlation function must take into account the behaviour of the underlying process we want to model. If the underlying phenomenon is continuously differentiable, the correlation function will likely show a parabolic behaviour near the origin, justifying the use of a Gaussian model. Conversely, physical phenomena usually show a linear behaviour near the origin, and Exponential or Spherical would usually perform better. Also note that for large distances the correlation is 0 according to the Spherical function for instance, while it is asymptotically 0 when applying other functions. Hence, another disadvantage of the parametric approach is that the choice of the model for the covariance function completely determines the convexity of the covariance or the regularity of the process characterized by the behaviour at the origin of the covariance function. Furthermore, a major drawback of parametric models is that the true covariance function may fail to belong to any of the considered parametric families.

To overcome the drawbacks of parametric methods, nonparametric procedures have received a growing attention along the last decades (see the works of Shapiro and Botha [85], Sampson and Guttorp [80], Hall et al. [47], Perrin et al. [75], Guillot et al. [46], Bel [10], Elogne [39], etc). Nonparametric approaches provide more flexibility when constructing the estimator of the covariance function, but their main problem comes from the difficulty to restrict to non-negative definite class of estimators. In the existent literature, some nonparametric estimators like the ones in Shapiro and Botha [85] and Hall et al. [47] verify this property, but in some cases the procedures are more complex, or their rates of convergence are not yet determined (see for example Shapiro and Botha [85], Sampson and Guttorp [80], Guillot et al. [46] and Bel [10]).

In this thesis, we will concentrate in the development of nonparametric estimators of the covariance function of a stochastic process which satisfy the non-negative definiteness property. We will study the theoretical properties of the estimators, and show their feasibility and practical performance through numerical simulations.

1.1.4 High dimensional setting

The covariance matrix $\mathbf{\Sigma}$ of a stochastic process $X = \left\{X\left(t\right) : t \in T \subset \mathbb{R}^d\right\}$ observed at the points $t_1, ..., t_n$ is defined by $\mathbf{\Sigma} = \left(Cov\left(X\left(t_j\right), X\left(t_k\right)\right)\right)_{1 \leq j,k \leq n}$.

The estimation of large covariance matrices, particularly in situations where the data dimension is comparable to or larger than the sample size, has attracted a lot of attention recently. The abundance of high-dimensional data is one reason for the interest in this problem: gene arrays, various kinds of spectroscopy, climate studies, and many other applications often generate very high dimensions and moderate sample sizes. Another reason is the ubiquity of the covariance matrix in data analysis tools. Principal component analysis (PCA) and linear and quadratic discriminant analysis (LDA and QDA) require an estimate of the covariance matrix. Furthermore, recent advances in random matrix theory (see Johnstone [49] and Paul [74] for a review) allowed in-depth theoretical studies of the traditional estimator, the sample (empirical) covariance matrix, and showed that without regularization the sample covariance performs poorly in high dimensions. These results stimulate research on alternative estimators in high dimensions.

In this thesis, we will propose a penalized estimation method for the covariance matrix of a stochastic process in a high dimensional setting. We will study the theoretical properties of the estimators, and show their practical behaviour through numerical examples.

1.2 Contributions of the thesis

In this section, we describe the nonparametric covariance estimation procedures proposed in this work. First, we give a general presentation of the chapters. Afterwards, we set the theoretical background in which our methods are based, and the different covariance estimators are introduced.

1.2.1 Chapters presentation

Our work falls into the following chapters.

Chapter 1 is divided into two parts. In the first part we present the underlying issue at the core of this thesis, which motivates the search of nonparametric estimators of the covariance function of a stochastic process under different conditions on the process. We give some definitions related to random fields such as stationarity and isotropy. Afterwards, we recall some characterization properties and the spectral representation of stationary covariance functions. Classical methods for covariance estimation are also briefly described. In the second part we give fundamental notions of wavelet thresholding, model selection and ℓ_1-type regularization techniques, which are the main procedures that allow us to define the nonparametric covariance estimators proposed in the next three chapters, respectively. We also present a short description of the estimators.

In **Chapter 2** we study the problem of nonparametric adaptive estimation of the covariance function of a stationary Gaussian process. For this purpose, we consider a wavelet-based method which combines the ideas of wavelet approximation and estimation by information projection in order to warrants the non-negative definiteness property of the solution. The spectral density of the process is estimated by projecting the wavelet thresholding expansion of the periodogram onto a family of exponential functions. This ensures that the spectral density estimator is a strictly positive function. Then, by Bochner's theorem, we obtain a non-negative definite estimator of the covariance function. The theoretical behavior of the estimator is established in terms of rate of convergence of the Kullback-Leibler discrepancy over Besov classes. We also show the good practical performance of the estimator through numerical experiments.

In **Chapter 3** we propose a model selection approach for nonparametric covariance estimation of nonstationary stochastic processes. Under very general assumptions, observing independent and identically distributed replications of the process at fixed observation points, we construct an estimator of the covariance function by expanding the process onto a collection of basis functions. This is based on the formulation of the covariance estimation problem through linear matrix regression models. We study non asymptotic, finite sample properties of this estimate and give a tractable way of selecting the best estimator among a possible set of candidates. The optimality of the procedure is proved via an oracle inequality which warrants that the best model is selected. Some numerical experiments show the good performance of the estimator proposed. The contents of this chapter is being published in *Electronic Journal of Statistics* (Bigot, Biscay, Loubes and Muñiz [14]).

Finally, in **Chapter 4** we present the Group-Lasso estimator of the covariance matrix of a stochastic process. We propose to estimate the covariance matrix in a high-dimensional setting under the assumption that the process has a sparse representation in a large dictionary of basis functions. Using a matrix regression model, we propose a new methodology for high-dimensional covariance matrix estimation based on empirical contrast regularization by a Group-Lasso penalty. Using such a penalty, the method selects a sparse set of basis functions in the dictionary used to approximate the process, leading to an approximation of the covariance matrix into a low dimensional space. Hence, we propose a new method of dimension reduction for high-dimensional data. Consistency of the estimator is studied in Frobenius and operator norms and an application to sparse Principal Components Analysis (PCA) is proposed.

1.2.2 Description of the estimation procedures

The covariance estimators introduced in this thesis are based on different nonparametric techniques: wavelet thresholding and information projection (Chapter 2), model selection (Chapter 3), and Group-Lasso regularization (Chapter 4). We present here the basic concepts of these methods, and briefly decribed the estimators that we proposed.

Estimation procedure based on wavelet thresholding

The central idea of the wavelet theory is the representation and analysis of data according to simultaneous location in the space and frequency domains (see the book of Mallat [63]). This idea can be traced back to the beginning of the last century, when Haar presented the first known wavelet system. We summarize now its fundamentals elements using current terminology.

Haar system and wavelet decomposition. One defines a father function

$$\phi(x) := \begin{cases} 1 & x \in [0,1) \\ 0 & \text{otherwise} \end{cases},$$

whose translations and scalings

$$\phi_{j,k}(x) := 2^{j/2}\phi(2^j x - k),\ j \geq 0,\ k = 0, ..., 2^j - 1,$$

called scaling functions, induce a multiresolution analysis of $L_2([0,1])$. This means that the scale of linear spaces $\{V_j\}_{j\geq 0}$, each defined by

$$V_j := span\{\phi_{j,k} : k = 0, ..., 2^j - 1\},$$

are nested, i.e. $V_0 \subset V_1 \subset ... \subset V_j \subset ... \subset L_2([0,1])$ and their union is dense in $L_2([0,1])$. Now, one defines the so called mother function as the particular linear combination of scaling functions

$$\psi\left(x\right) := \frac{1}{\sqrt{2}}\phi_{1,0}(x) - \frac{1}{\sqrt{2}}\phi_{1,1}(x),$$

which yields $\int_0^1 \phi(x)\psi\left(x\right)dx = 1$. The translations and scalings

$$\psi_{j,k}(x) := 2^{j/2}\psi(2^j x - k),\ j \geq 0,\ k = 0, ..., 2^j - 1$$

of the mother function are called wavelets. Rescaling $\psi\left(x\right)$, one easily sees that the spans of all the wavelets on each dyadic level,

$$W_j := span\{\psi_{j,k} : k = 0, ..., 2^j - 1\},$$

constitute orthogonal complements to the spaces V_j, that is, $V_{j+1} = V_j \oplus Wj$. We have an orthonormal basis for $L_2([0,1])$. Choosing a coarsest level $j_0 \geq 0$, every function $g \in L_2([0,1])$ has an unique wavelet expansion of the form

$$g\left(\omega\right) = \sum_{k=0}^{2^{j_0}-1} \left\langle g, \phi_{j_0,k}\right\rangle \phi_{j_0,k}\left(\omega\right) + \sum_{j=j_0}^{+\infty}\sum_{k=0}^{2^j-1} \left\langle g, \psi_{j,k}\right\rangle \psi_{j,k}\left(\omega\right),$$

where the expansion coefficients fulfill

$$\left\langle g, \phi_{j_0,k}\right\rangle = \int_0^1 g(x)\phi_{j_0,k}\left(x\right)dx \text{ and } \left\langle g, \psi_{j,k}\right\rangle = \int_0^1 g(x)\psi_{j,k}\left(x\right)dx.$$

Although the main ideas are already present in the simple Haar system, its poor regularity has limited its applications. The real breakthrough of wavelets starts in the late 80's, as the works of Mallat [63] and Daubechies [31] showed the way to construct wavelet families satisfying more demanding requirements on properties as symmetry, orthogonality, compact support, smoothness and vanishing moments (a wavelet is said to have n vanishing moments if $\int_0^1 \psi\left(x\right)x^m dx$ is identically zero for $m = 0, ..., n-1$ but not for $m = n$). In particular, Daubechies [31] presents a family of orthonormal wavelets in $L_2(\mathbb{R})$ with compact support and arbitrary regularity, yielding a powerful tool to represent the amount of information conveyed by functional data.

Wavelet thresholding. Wavelet thresholding is a nonlinear technique that was developed in the early 1990's (see Donoho and Johnstone [35]) and has been used in a wide range of applications including meteorology (Briggs and Levine [22] and Katul and Vidakovic [53]). The key element of this procedure is the modification of the wavelet coefficients, which is supposed to separate deterministically predictable features from random error. Coefficients that are suspected to have a high signal-to-noise ratio are retained and coefficients with a presumably low signal-to-noise ratio are curtailed in some way. More specifically, wavelet thresholding methods use the absolute value of the wavelet coefficient as a criterion in the curtailing step. This is motivated by the fact that the measures of smoothness of a function often depend on the magnitudes of its wavelet coefficients.

The two most common thresholding strategies are the hard and soft thresholding rules (see Donoho and Johnstone [35]). Hard thresholding retains wavelet coefficients that possess an absolute value greater than or equal to a universal threshold ξ and sets others to zero, while soft thresholding additionally reduces the absolute values of the retained coefficients by the value of ξ.

Level-dependent thresholds. The procedure of wavelet thresholding can be extended from using a universal threshold ξ to using level-dependent thresholds ξ_j. Statistical arguments sustain that the largest scale's wavelet coefficients should be left un-thresholded regardless of their size (Donoho and Johnstone [35]). Physical arguments state that signal characteristics differ at different levels, because coarse levels show a larger proportion of important signal features. Even if noise is the same on each level, smaller thresholds on coarse levels should be preferred.

Statistical optimality properties. In addition to the intuitive arguments above, several statistical optimality properties have been proven for the wavelet thresholding approach (see Donoho and Johnstone [35], Donoho et al. [36], Härdle et al. [48]). Under appropriate regularity conditions, if the threshold is optimally chosen then the estimator attains smoothness and adaptation.

Smoothness states that there is a high probability that the estimator is as smooth as the unknown function to be estimated (see Vidakovic [89]). Adaptation refers to the so-called minimax risk of the estimator. The risk is the expected distance between the estimator and the unknown function, where the distance is determined by a norm which can be quite general. The minimax risk minimizes the worst case risk of the estimator within the function space. Adaptation states that the estimator achieves almost minimax risk over a wide range of smoothness classes, including the classes in which linear estimators do not achieve the minimax rate (Vidakovic [89]).

In Chapter 2 we use a hard thresholding technique with level-dependent thresholds to tackle the problem of adaptive estimation of the spectral density and the corresponding covariance function of a stationary Gaussian process. In the next section we give a short introduction to the wavelet-based estimators that we propose.

Description of the estimators

Consider the estimation of the spectral density f of a stationary Gaussian process $(X_t)_{t\in\mathbb{N}}$ observed at n regularly spaced points. Let

$$I_n(\omega) = \frac{1}{2\pi n}\sum_{t=1}^{n}\sum_{t'=1}^{n}\left(X_t - \overline{X}\right)\left(X_{t'} - \overline{X}\right)^* e^{-i2\pi\omega(t-t')},\ \omega \in [0,1]$$

be the classical periodogram, where $X_1, ..., X_n$ is the observed sample, $\left(X_t - \overline{X}\right)^*$ denotes the conjugate transpose of $\left(X_t - \overline{X}\right)$ and $\overline{X} = \frac{1}{n}\sum\limits_{t=1}^{n} X_t$. The expansion of $I_n(\omega)$ onto a wavelet basis allows to obtain estimators of the wavelet coefficients $a_{j_0,k}$ and $b_{j,k}$ of f

given by

$$\widehat{a}_{j_0,k} = \int_0^1 I_n(\omega)\,\phi_{j_0,k}(\omega)\,d\omega \quad \text{and} \quad \widehat{b}_{j,k} = \int_0^1 I_n(\omega)\,\psi_{j,k}(\omega)\,d\omega.$$

To simplify the notations, we write $(\psi_{j,k})_{j=j_0-1}$ for the scaling functions $(\phi_{j,k})_{j=j_0}$ and we denote by $\widehat{\beta}_{j,k}$ both the estimation of the scaling coefficients $\widehat{a}_{j_0,k}$ and the wavelet coefficients $\widehat{b}_{j,k}$. We consider the hard thresholding rule defined by

$$\delta_{\xi_{j,n}}\left(\widehat{\beta}_{j,k}\right) = \widehat{\beta}_{j,k} I\left(|\widehat{\beta}_{j,k}| \geq \xi_{j,n}\right),$$

where $\xi_{j,n}$ is an appropriate level-dependent threshold. For $j_1 \geq j_0$ we define

$$\Lambda_{j_1} = \left\{(j,k) : j_0 - 1 \leq j < j_1, 0 \leq k \leq 2^j - 1\right\}.$$

Let θ denotes a vector in $\mathbb{R}^{|\Lambda_{j_1}|}$, where $|\Lambda_{j_1}|$ denotes the cardinality of Λ_{j_1}. The wavelet-based exponential family $\mathfrak{E}_{j_1}$ at scale j_1 is defined as the set of functions:

$$\mathfrak{E}_{j_1} = \left\{ f_{j_1,\theta}(.) = \exp\left(\sum_{(j,k)\in\Lambda_{j_1}} \theta_{j,k}\psi_{j,k}(.) \right),\ \theta = (\theta_{j,k})_{(j,k)\in\Lambda_{j_1}} \in \mathbb{R}^{|\Lambda_{j_1}|} \right\}.$$

The spectral density f is estimated by projecting the wavelet thresholding expansion of the periodogram onto the family $\mathfrak{E}_{j_1}$ of exponential functions. More specifically, we estimate f by searching for some $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$ such that

$$\left\langle f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}, \psi_{j,k} \right\rangle = \delta_{\xi_{j,n}}\left(\widehat{\beta}_{j,k}\right) \text{ for all } (j,k) \in \Lambda_{j_1},$$

where $f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}} \in \mathfrak{E}_{j_1}$. The resulting nonlinear estimator $f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}$ is a strictly positive function by construction. Therefore, the corresponding estimator of the covariance function $\widehat{\sigma}$ (which is obtained as the inverse Fourier transform of $f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}$) is a non-negative definite function.

Estimation procedure based on model selection

Model selection is a classical topic in statistics (see the book of Massart [65]). The idea of selecting a model via penalizing an empirical criterion goes back to the early seventies with the pioneering works of Mallows [64] and Akaike [2]. One can find many consistency results in the literature for such criteria. These results are asymptotic in the sense that one deals with a given number of models and the number of observations tends to infinity. We shall give an overview of a non asymptotic theory for model selection which has emerged during these last fifteen years.

Model selection. If one observes some random variable X (which can be a random vector or a random process) with unknown distribution, the basic problem of statistical inference is to take a decision about some quantity σ related to the distribution of X; for instance to give an estimate of σ or provide a confidence set for σ with a given level of confidence. Usually, one starts from a genuine estimation procedure for σ and try to get some idea of how far it is from the target. Since generally speaking the exact distribution of the estimation procedure is not available, the role of Probability Theory is to provide relevant approximation tools to evaluate it.

Designing a genuine estimation procedure requires some prior knowledge on the unknown distribution of X and choosing a proper model is a major problem for the statistician. The aim of model selection is to construct data-driven criteria to select a model among a given list. In many situations, motivated by applications such as signal analysis, it is useful to allow the size of the models to depend on the size N of the sample $X_1, ..., X_N$ of X. In these situations, classical asymptotic analysis breaks down and one needs to introduce an alternative non asymptotic approach. By non asymptotic, we do not mean of course that large samples of observations are not taken into account but that the size of the models as well as the size of the list of models should be allowed to be large when N is large in order to be able to warrant that the statistical model is not far from the truth. When the target quantity σ to be estimated is a function, this allows in particular to consider models which have good approximation properties at different scales and use model selection criteria to choose from the data the best approximating model. One of the most commonly used method to estimate σ is minimum contrast estimation.

Minimum contrast estimation. One can typically think of σ as a function belonging to some space $\mathcal{S}$ which may be infinite dimensional. Consider some empirical criterion L_N, i.e. based on the observations $X_1, ..., X_N$, such that on the set $\mathcal{S}$, $\upsilon \rightarrow \mathbb{E}\left[L_N\left(\upsilon\right)\right]$ achieves a minimum at the point σ. Such a criterion is called an empirical contrast for the estimation of σ. Given some subset S of $\mathcal{S}$ (that we call a model), a minimum contrast estimator $\widehat{\sigma}$ of σ is a minimizer of L_N over S. The underlying idea in minimum contrast estimation is that, if one substitutes the empirical criterion L_N to its expectation and minimizes L_N on S, there is some hope to get a sensible estimator of σ, at least if σ belongs (or is close enough) to the model S. A given empirical criterion is indeed an empirical contrast if the associated natural loss function $l(\sigma, \upsilon) = \mathbb{E}\left[L_N\left(\upsilon\right)\right] - \mathbb{E}\left[L_N\left(\sigma\right)\right]$ is non-negative for all $\upsilon \in S$.

The model choice paradigm. The main problem which arises from minimum contrast estimation in a parametric setting is the choice of a proper model S on which the minimum contrast estimator is to be defined. In other words, it may be difficult to guess what is the right parametric model to consider in order to reflect the nature of data from the real life and one can get into problems whenever the model S is false in the sense that the true σ is too far from S. One could then be tempted to choose S as big as possible. Taking S as $\mathcal{S}$ itself or as a "huge" subset of $\mathcal{S}$ is known to lead to inconsistent (see Bahadur [6]) or suboptimal estimators (see Birgé and Massart [16]). We see that choosing some model S in advance leads to some difficulties:

- If S is a small model (think of some parametric model, defined by 1 or 2 parameters for instance) the behavior of a minimum contrast estimator on S is satisfactory as long as σ is close enough to S but the model can easily turn to be false.

- On the contrary, if S is a huge model (think of the set of all continuous functions on $[0,1]$ in the regression framework for instance), the minimization of the empirical criterion leads to a very poor estimator of σ even if σ truly belongs to S.

It is therefore interesting to consider a family of models instead of a single one and try to select some appropriate model among the family. More precisely, if we consider some empirical contrast L_N and some (at most countable and usually finite) collection of models $(S_m)_{m\in\mathcal{M}}$, let us represent each model S_m by the minimum contrast estimator $\widehat{\sigma}_m$ related to L_N. The purpose is to select the best estimator among the collection $(\widehat{\sigma}_m)_{m\in\mathcal{M}}$. Ideally, one would like to consider $m(\sigma)$ minimizing the risk $\mathbb{E}\left[l(\sigma,\widehat{\sigma}_m)\right]$ with respect to $m\in\mathcal{M}$. The minimum contrast estimator $\widehat{\sigma}_{m(\sigma)}$ on the corresponding model $S_{m(\sigma)}$ is called an oracle (according to the terminology introduced by Donoho and Johnstone, see [35] for instance). Unfortunately, since the risk depends on the unknown parameter σ, so does $m(\sigma)$ and the oracle is not an estimator of σ. However, the risk of an oracle can serve as a benchmark which will be useful in order to evaluate the performance of any data driven selection procedure among the collection of estimators $(\widehat{\sigma}_m)_{m\in\mathcal{M}}$. Note that this notion is different from the notion of true model. In other words if σ belongs to some model S_{m_0}, this does not necessarily imply that $\widehat{\sigma}_{m_0}$ is an oracle. The idea is now to consider data-driven criteria to select an estimator which tends to mimic an oracle, i.e. one would like the risk of the selected estimator $\widehat{\sigma}_{\widehat{m}}$ to be as close as possible to the risk of an oracle.

Model selection via penalization. The model selection via penalization procedure consists in considering some proper penalty function $pen:\mathcal{M}\to\mathbb{R}_+$ and take $\widehat{m}$ minimizing the penalized criterion

$$L_N\left(\widehat{\sigma}_m\right)+pen(m)$$

over $\mathcal{M}$. We can then define the selected model $S_{\widehat{m}}$ and the corresponding estimator $\widehat{\sigma}_{\widehat{m}}$.

This method has many applications in Statistics. Penalized criteria have been proposed by Akaike [2] for penalized log-likelihood in the density estimation framework and by Mallows [64] for penalized least squares regression, where the variance of the errors of the regression framework is assumed to be known for the sake of simplicity. In both cases the penalty functions are proportional to the number of parameters D_m of the corresponding model S_m. Penalties can also aim at imposing some smoothness constraint, in particular ℓ_1 penalties is considered for instance in this framework in Loubes and van de Geer [59].

Oracle inequalities. The performance of the penalized estimator $\widehat{\sigma}_{\widehat{m}}$ is usually studied via non asymptotic risk bounds, which express that it performs almost as well as if the best model (i.e. with minimal risk) were known. This methodology heavily relies on concentration inequalities, which allow to obtain oracle inequalities of the form

$$\mathbb{E}\left[l(\sigma,\widehat{\sigma}_{\widehat{m}})\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}\left[l(\sigma,\widehat{\sigma}_{m})\right] + O\left(\frac{1}{N}\right),$$

which can be interpreted by saying that the risk of the selected estimator $\widehat{\sigma}_{\widehat{m}}$ is as close as possible to the risk of the oracle.

In Chapter 3 we use a model selection procedure to construct a penalized least squares estimator of the covariance function of a stochastic process. A key idea in this approach is to state the covariance function estimation problem by means of matrix regression models based on function expansions of the process. We obtain oracle inequalities which warrant that the selection procedure is optimal in the sense that it behaves as if the true model were at hand. In the next section we briefly explain the main ideas of the covariance estimation procedure by model selection that we propose.

Description of the estimators

Consider a zero mean stochastic process $\left\{X\left(t\right) : t \in T \subset \mathbb{R}^{d},\ d \in \mathbb{N}\right\}$ with finite covariance function, i.e. $\left|\sigma\left(s,t\right)\right| = \left|Cov\left(X\left(s\right),X\left(t\right)\right)\right| < +\infty$ for all $s,t \in T$. The observations are $X_i\left(t_j\right)$ for $i = 1,...,N$, $j = 1,...,n$, where the points $t_1,...,t_n \in T$ are fixed, and $X_1,...,X_N$ are independent copies of the process X. To estimate the covariance function σ we consider a functional expansion $\widetilde{X}$ to approximate the underlying process X and take the covariance of $\widetilde{X}$ as an approximation of the true covariance σ. For this, let $\left\{g_\lambda\right\}_\lambda$ be a collection of possibly independent functions $g_\lambda : T \to \mathbb{R}$, and define $\mathcal{M}$ as a generic countable set given by $\mathcal{M} = \left\{m : m \text{ is a set of indices of cardinality } \left|m\right|\right\}$. Set $m \in \mathcal{M}$, an approximation to the process X is of the following form:

$$\widetilde{X}\left(t\right) = \sum_{\lambda\in m} a_\lambda g_\lambda\left(t\right),$$

where a_λ, with $\lambda \in m$, are suitable random coefficients. It is natural to consider the covariance function ρ of $\widetilde{X}$ as an approximation of σ. The covariance ρ can be written as $\rho\left(s,t\right) = \mathbf{G}_s^\top\overline{\mathbf{\Psi}}\mathbf{G}_t$, where, after reindexing the functions if necessary, $\mathbf{G}_t = \left(g_\lambda\left(t\right),\lambda \in m\right)^\top$ and $\overline{\mathbf{\Psi}} = \left(\mathbb{E}\left(a_\lambda a_\mu\right)\right)$ with $\left(\lambda,\mu\right) \in m\times m$. Hence we are led to look for an estimate $\widehat{\sigma}$ of σ in the class of functions of the form $\mathbf{G}_s^\top\mathbf{\Psi}\mathbf{G}_t$, where $\mathbf{\Psi} \in \mathbb{R}^{|m|\times|m|}$ is some symmetric matrix. Denote by $\mathbf{\Sigma} = \left(\sigma\left(t_j,t_k\right)\right)_{1\leq j,k\leq n}$ the true covariance matrix while $\mathbf{\Gamma} = \left(\rho\left(t_j,t_k\right)\right)_{1\leq j,k\leq n}$ denotes the covariance matrix of the approximated process $\widetilde{X}$ at the observation points. Hence $\mathbf{\Gamma} = \mathbf{G}\overline{\mathbf{\Psi}}\mathbf{G}^\top$, where $\mathbf{G}$ is the $n \times |m|$ matrix with entries $g_{j\lambda} = g_\lambda\left(t_j\right)$, $j = 1,...,n$, $\lambda \in m$. $\mathbf{G}$ will be called the design matrix corresponding to the set of basis functions indexed by m. Comparing the covariance function ρ with the true one σ over the design points t_j, implies quantifying the deviation of $\mathbf{\Gamma}$ from $\mathbf{\Sigma}$. For this consider the following loss function

$$L\left(\mathbf{\Psi}\right) = \mathbb{E}\left\|\mathbf{x}\mathbf{x}^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2,$$

where $\mathbf{x} = \left(X\left(t_1\right),...,X\left(t_n\right)\right)^\top$ and $\left\|\mathbf{A}\right\|_F = \sqrt{\mathrm{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)}$ is the Frobenius matrix norm defined for all matrix $\mathbf{A}$ with real entries. Note that

$$L\left(\mathbf{\Psi}\right) = \left\|\mathbf{\Sigma} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + C,$$

where the constant C does not depend on $\mathbf{\Psi}$. To the loss L corresponds the following empirical contrast function L_N, which will be the fitting criterion we will try to minimize

$$L_N(\mathbf{\Psi}) = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{G} \mathbf{\Psi} \mathbf{G}^\top \right\|_F^2 .$$

This loss is exactly the sum of the squares of the residuals corresponding to the matrix linear regression model

$$\mathbf{x}_i \mathbf{x}_i^\top = \mathbf{G} \mathbf{\Psi} \mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, ..., N,$$

with i.i.d. matrix errors $\mathbf{U}_i$, such that $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$. This remark provides a natural framework to study the covariance estimation problem as a matrix regression model. Note also that the set of matrices $\mathbf{G} \mathbf{\Psi} \mathbf{G}^\top$ is a linear subspace of $\mathbb{R}^{n \times n}$ when $\mathbf{\Psi}$ ranges over the space of symmetric matrices denoted by $\mathcal{S}_{|m|}$.

We propose the following minimum contrast estimation procedure: for a given design matrix $\mathbf{G}$, we define

$$\widehat{\mathbf{\Psi}} = \underset{\mathbf{\Psi} \in \mathcal{S}_{|m|}}{\arg\min} L_N(\mathbf{\Psi}),$$

and take $\widehat{\mathbf{\Sigma}} = \mathbf{G} \widehat{\mathbf{\Psi}} \mathbf{G}^\top$ as the least squares estimator of $\mathbf{\Sigma}$. It will be shown that $\widehat{\mathbf{\Psi}}$ is a non-negative definite matrix, thus $\widehat{\mathbf{\Sigma}}$ also has this property. Hence, the resulting estimator of the covariance function, given by $\widehat{\sigma}(s,t) = \mathbf{G}_s^\top \widehat{\mathbf{\Psi}} \mathbf{G}_t$, is a non-negative definite function.

The role of $\mathbf{G}$ and therefore the choice of the subset of indices m is crucial since it determines the behavior of the estimator. Hence, we aim at selecting the best design matrix $\mathbf{G} = \mathbf{G}_m$ among a collection of candidates $\left\{ \mathbf{G}_m \in \mathbb{R}^{n \times |m|},\ m \in \mathcal{M} \right\}$. Let $\widehat{\mathbf{\Sigma}}_m$ be the least squares covariance estimators corresponding to the matrix $\mathbf{G}_m$, with $m \in \mathcal{M}$. The problem of interest is to select the best of these estimators in the sense of the minimal quadratic risk given by

$$\mathbb{E} \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m \right\|_F^2 = \left\| \mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m \right\|_F^2 + \frac{\mathrm{Tr}\left((\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m) \mathbf{\Phi} \right)}{N},$$

where $\mathbf{\Pi}_m = \mathbf{G}_m \left(\mathbf{G}_m^\top \mathbf{G}_m \right)^- \mathbf{G}_m^\top$ with $m \in \mathcal{M}$ and $\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1 \mathbf{x}_1^\top \right) \right)$ (for any vector $\mathbf{z} = (Z_i)_{1 \le i \le n^2}$ the matrix $\mathbb{V}(\mathbf{z}) = (Cov(Z_i, Z_j))_{1 \le i,j \le n^2}$). For this, we use a model selection via penalization technique. We define the penalized covariance estimator $\widehat{\mathbf{\Sigma}}_{\widehat{m}}$ by

$$\widehat{m} = \underset{m \in \mathcal{M}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i \mathbf{x}_i^\top - \widehat{\mathbf{\Sigma}}_m \right\|_F^2 + pen(m) \right\},$$

where the penalty function $pen : \mathcal{M} \to \mathbb{R}$ is of the form

$$pen(m) = (1 + \theta) \frac{\mathrm{Tr}\left((\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m) \mathbf{\Phi} \right)}{N},$$

with $\theta > 0$.

The optimality of the model selection procedure is proved via an oracle inequality which states that the quadratic risk $\mathbb{E} \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{\widehat{m}} \right\|_F^2$ is bounded by the quadratic risk of the oracle, given by $\inf_{m \in \mathcal{M}} \mathbb{E} \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m \right\|_F^2$, except for a constant factor and an additive term of order $O\left(\frac{1}{N} \right)$.

Estimation procedure based on ℓ_1-type regularization techniques

During the last few years, a great deal of attention has been focused on ℓ_1-type regularization methods for the estimation of parameters in high dimensional linear regression when the number of variables can be much larger than the sample size. The most popular of these methods is the Lasso estimation procedure, proposed by Tibshirani [87]. The Lasso enjoys two important properties. First, it is naturally sparse, i.e., it has a large number of zero components. Second, it is computationally feasible even for high-dimensional data (Efron et al. [37] and Osborne et al. [73]). In the literature, the theoretical and empirical properties of the Lasso procedure have been extensively studied. See, for instance, Efron et al. [37], Fu and Knight [43], Meinshausen and Buhlmann [68], Tibshirani [87], Van der Geer [88], Zhang and Huang [91], and Zhao and Yu [92] for instance). Recent extensions of the Lasso and their performances can be found in Bickel, Ritov and Tsybakov [13] and Lounici [61].

Lasso estimators have been also studied in the nonparametric regression setting (see Bunea, Tsybakov and Wegkamp [24], [25], Greenshtein and Ritov [45] and Nemirovski [71]). In particular, Bunea, Tsybakov and Wegkamp [24], [25] obtain sparsity oracle inequalities for the prediction loss in this context and point out the implications for minimax estimation in classical nonparametric regression settings, as well as for the problem of aggregation of estimators.

Lasso estimator. Consider the problem of recovering a sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^q$ using a sample of independent pairs $(\mathbf{A}_{1\cdot}; Y_1), ..., (\mathbf{A}_{n\cdot}; Y_n)$ from a multiple linear regression model,

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{1.5}$$

where $\mathbf{Y}$ is the $n \times 1$ response vector, $\mathbf{A}$ represents the observed $n \times q$ design matrix whose i-th row vector is denoted by $\mathbf{A}_{i\cdot}$, $\boldsymbol{\beta}^*$ is the true unknown coefficient vector that we want to recover, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$ is an $n \times 1$ zero-mean random vector such that $\mathbb{E}\left(\varepsilon_i^2\right) < +\infty$, $1 \leq i \leq n$. The Lasso estimator $\widehat{\boldsymbol{\beta}}_L$ solves the minimization problem

$$\widehat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta} \in \mathbb{R}^q}{\arg\min} \left\{ \|\mathbf{Y} = \mathbf{A}\boldsymbol{\beta}\|_{\ell_2}^2 + 2\lambda \|\boldsymbol{\beta}\|_{\ell_1} \right\},$$

where $\lambda > 0$ is called the regularization parameter and the penalty term is such that give preference to solutions with components $\beta_j = 0$, where β_j, $1 \leq j \leq q$, denotes the j-th component of $\boldsymbol{\beta}$.

Group-Lasso estimator. The Group-Lasso estimator, proposed by Yuan and Lin [90], is an extension of the Lasso estimator. It can be described as follows. We are interested in the situation where all the variables in model (1.5) are naturally partitioned into M groups, where M is fixed and finite. Suppose the number of variables in the k-th group is d_k, and thus by definition we have $q = \sum\limits_{k=1}^{M} d_k$. We can rewrite the linear model (1.5) as

$$\mathbf{Y} = \sum_{k=1}^{M} \mathbf{A}_k \boldsymbol{\beta}_k^* + \boldsymbol{\varepsilon},$$

where $\mathbf{A}_k$ is an $n \times d_k$ matrix corresponding to the k-th group and $\boldsymbol{\beta}_k^*$ is the corresponding $d_k \times 1$ coefficient subvector of $\boldsymbol{\beta}^*$. The Group-Lasso estimator is defined as the solution of the following optimization problem:

$$\widehat{\boldsymbol{\beta}}_{GL} = \underset{\beta \in \mathbb{R}^q}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{A}\boldsymbol{\beta}\|_{\ell_2}^2 + 2\lambda \sum_{k=1}^{M} \gamma_k \|\boldsymbol{\beta}_k\|_{\ell_2} \right\}, \tag{1.6}$$

where λ is a positive number which penalizes complex model, and $\gamma_k > 0$ is multiplied over each group. The Group-Lasso can be viewed as a generalization of the Lasso for the grouped variables by replacing the ℓ_1-regularization with the sum of ℓ_2-norm regularization. The penalty term imposes to give preference to solutions with components $\boldsymbol{\beta}_k = \mathbf{0}$, where $\boldsymbol{\beta}_k$, $1 \leq k \leq M$, denotes the $d_k \times 1$ coefficient subvector of $\boldsymbol{\beta}$.

In Chapter 4 we propose the Group-Lasso estimator of the covariance matrix of a stochastic process in a high-dimensional setting under the assumption that the process has a sparse representation in a large dictionary of basis functions. We extend the definition of the Group-Lasso estimator (1.6) to the case of a matrix regression model and give consistency results using oracle inequalities. In the next section we briefly explain the main ideas of the estimation procedure by Group-Lasso regularization that we propose.

Description of the estimators

Let T be some subset of $\mathbb{R}^d$, $d \in \mathbb{N}$, and let $X = \{X(t), t \in T\}$ be a zero mean stochastic process with values in $\mathbb{R}$ and finite covariance function $\sigma(s,t) = \mathbb{E}(X(s)X(t))$ for all $s,t \in T$. Let $t_1, \ldots, t_n$ be fixed points in T (deterministic design), $X_1, ..., X_N$ independent copies of the process X, and suppose that we observe the noisy processes

$$\widetilde{X}_i(t_j) = X_i(t_j) + \mathcal{E}_i(t_j) \text{ for } i = 1, ..., N, \; j = 1, ..., n, \tag{1.7}$$

where $\mathcal{E}_1, ..., \mathcal{E}_N$ are independent copies of a second order Gaussian process $\mathcal{E}$ with zero mean and independent of X, which represent an additive source of noise in the measurements. We consider the problem of the estimation of the covariance matrix $\boldsymbol{\Sigma}$ of the process X at the design points in a high-dimensional setting, i.e., when $n >> N$ or $n \sim N$ from the noisy observations (1.7). We suppose that the process X has a sparse representation in a large dictionary of basis functions, that is

$$X(t) \approx \sum_{m=1}^{M} a_m g_m(t), \tag{1.8}$$

where $g_m : T \to \mathbb{R}$, $m = 1, ..., M$, denote the basis functions in the dictionary, a_m, $m = 1, ..., M$ are real valued random variables and the notation $\approx$ means that the process X can be well approximated by the functions in the dictionary. Hence, for each trajectory X_i,

$$X_i(t_j) \approx \sum_{m=1}^{M} a_{i,m} g_m(t_j), \; i = 1, ..., N, \; j = 1, ..., n, \tag{1.9}$$

where $a_{i,m}$, $m = 1, ..., M$, $i = 1, ..., N$ are real valued random variables. Then (1.9) can be written in matrix notation as $\mathbf{X}_i \approx \mathbf{G}\mathbf{a}_i$, $i = 1, ..., N$, where $\mathbf{G}$ is the $n \times M$ matrix with entries $\mathbf{G}_{jm} = g_m(t_j)$ with $1 \leq j \leq n$ and $1 \leq m \leq M$, and for $i = 1, ..., N$, we denote by $\mathbf{a}_i$ the $M \times 1$ random vector of components $a_{i,m}$, with $1 \leq m \leq M$. Since $\mathbf{X} \approx \mathbf{Ga}$ with $\mathbf{a} = (a_m)_{1 \leq m \leq M}$ and a_m as in (1.8), it follows that

$$\mathbf{\Sigma} \approx \mathbb{E}\left(\mathbf{Ga}(\mathbf{Ga})^\top\right) = \mathbb{E}\left(\mathbf{Gaa}^\top\mathbf{G}^\top\right) = \mathbf{G\Psi}^*\mathbf{G}^\top \text{ with } \mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{aa}^\top\right).$$

Denote by $\widetilde{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top$ the sample covariance matrix from the noisy observations (1.7) and consider the following matrix regression model

$$\widetilde{\mathbf{S}} = \mathbf{\Sigma} + \mathbf{U} + \mathbf{W}, \tag{1.10}$$

where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is a centered error matrix given by $\mathbf{U} = \mathbf{S} - \mathbf{\Sigma}$, and $\mathbf{W} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{W}_i$, where $\mathbf{W}_i = \mathcal{E}_i\mathcal{E}_i^\top \in \mathbb{R}^{n \times n}$ and $\mathcal{E}_i = \left(\mathcal{E}_i(t_1), ..., \mathcal{E}_i(t_n)\right)^\top$, $i = 1, \ldots, N$.

The size M of the dictionary can be very large, but it is expected that the process X has a sparse expansion in this basis, meaning that, in approximation (1.8), many of the random coefficients a_m are close to zero. We are interested in obtaining an estimate of the covariance $\mathbf{\Sigma}$ in the form $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ such that $\widehat{\mathbf{\Psi}}$ is a symmetric $M \times M$ matrix with many zero rows (and so, by symmetry, many corresponding zero columns). Note that setting the k-th row of $\widehat{\mathbf{\Psi}}$ to $\mathbf{0} \in \mathbb{R}^M$ means to remove the function g_k from the set of basis functions $(g_m)_{1 \leq m \leq M}$ in the function expansion associated to $\mathbf{G}$. To select a sparse set of rows/columns in the matrix $\widehat{\mathbf{\Psi}}$ we use a Group-Lasso approach to threshold some rows/columns of $\widehat{\mathbf{\Psi}}$. We define the Group-Lasso estimator of the covariance matrix $\mathbf{\Sigma}$ by $\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}_\lambda\mathbf{G}^\top \in \mathbb{R}^{n \times n}$, where $\widehat{\mathbf{\Psi}}_\lambda$ is the solution of the following optimization problem:

$$\widehat{\mathbf{\Psi}}_\lambda = \underset{\mathbf{\Psi} \in \mathcal{S}_M}{\arg\min}\left\{\left\|\widetilde{\mathbf{S}} - \mathbf{G\Psi G}^\top\right\|_F^2 + 2\lambda\sum_{k=1}^{M}\gamma_k\left\|\mathbf{\Psi}_k\right\|_{\ell_2}\right\},$$

where $\mathbf{\Psi}$ is an $M \times M$ symmetric matrix, λ is a positive regularization parameter and γ_k are some weights whose values will be discuss later on. Note that the penalty term imposes to give preference to solutions with components $\mathbf{\Psi}_k = \mathbf{0}$, where $(\mathbf{\Psi}_k)_{1 \leq k \leq M}$ denotes the columns of $\mathbf{\Psi}$. Thus $\widehat{\mathbf{\Psi}}_\lambda \in \mathbb{R}^{M \times M}$ can be interpreted as the Group-Lasso estimator of $\mathbf{\Sigma}$ in the matrix regression model (1.10).

Chapter 2

Adaptive estimation of covariance functions via wavelet thresholding and information projection

Abstract: In this chapter, we study the problem of nonparametric adaptive estimation of the covariance function of a stationary Gaussian process. For this purpose, we consider a wavelet-based method which combines the ideas of wavelet approximation and estimation by information projection in order to warrants the non-negative definiteness property of the solution. The spectral density of the process is estimated by projecting the wavelet thresholding expansion of the periodogram onto a family of exponential functions. This ensures that the spectral density estimator is a strictly positive function. Then, by Bochner's theorem, we obtain a non-negative definite estimator of the covariance function. The theoretical behavior of the estimator is established in terms of rate of convergence of the Kullback-Leibler discrepancy over Besov classes. We also show the good practical performance of the estimator in some numerical experiments.

2.1 Introduction

Estimating a covariance function is a fundamental problem in inference for stationary stochastic processes. Many applications in several fields such as geosciences, ecology, demography and financial time series are deeply related to this issue, see for instance Journel and Huijbregts [52], Christakos [26] and Stein [86]. The purpose of this work is not only to provide an estimator but also to guarantee that it is a covariance function. In particular, we aim at preserving the property of non-negative definiteness.

For this, usually statisticians resort to fitting parametric models, which are numerous in the literature, see Cressie [28] for a detailed account of parametric covariance estimation. Nonparametric approaches provide more flexibility when constructing the estimator but their main drawback comes from the difficulty to restrict to non-negative definite class of estimators. For example, Shapiro and Botha [85] suggest an estimator with this property on a discrete set but not on the continuum. In the nonstationary case and for spatio-temporal data, Sampson and Guttorp [80] propose an approach based on the particular covariance representation due to Schoenberg [83], which ensures that the resulting estimator is a covariance function. Hall, Fisher and Hoffman [47] enforce the non-negative definiteness property on a kernel type estimator of the covariance function using Bochner's theorem [20], which characterizes the class of continuous non-negative definite functions by the behavior of their Fourier transform. However, this approach requires the precise choice of three parameters, including an optimal selection of the bandwidth of the kernel function. More recently Elogne, Perrin and Thomas-Agnan [40] use interpolation methods for estimating smooth stationary covariance. The major drawback is that the computation of this estimator is difficult since it involves the calculation of convolution integrals.

Bochner's theorem states that a continuous function on $\mathbb{R}^d$ is non-negative definite if and only if it is the Fourier transform of a bounded non-negative measure called the spectral measure. When a spectral measure has a density, this density is called the spectral density. Hence, the estimation of the covariance function is strongly related to the estimation of the spectral density of the process.

Actually, inference in the spectral domain uses the periodogram of the data, providing an inconsistent estimator which must be smoothed in order to achieve consistency. For highly regular spectral densities, linear smoothing techniques such as kernel smoothing are appropriate (see Brillinger [23]). However, linear smoothing methods are not able to achieve the optimal mean-square rate of convergence for spectra whose smoothness is distributed inhomogeneously over the domain of interest. For this, nonlinear methods are needed. A nonlinear method for adaptive spectral density estimation of a stationary Gaussian sequence was proposed by Comte [27]. It is based on model selection techniques. Others nonlinear smoothing procedures are the wavelet thresholding methods, first proposed by Donoho and Johnstone [35]. In this context, different thresholding rules have been proposed by Neumann [72] and Fryzlewics, Nason and von Sachs [42] to name but a few.

Neumann's approach [72] consists in pre-estimating the variance of the periodogram via kernel smoothing, so that it can be supplied to the wavelet estimation procedure. Kernel pre-estimation may not be appropriate in cases where the underlying spectral density is of low regularity. One way to avoid this problem is proposed in Fryzlewics, Nason and von Sachs [42], where the empirical wavelet coefficient thresholds are built as appropriate local weighted ℓ_1 norms of the periodogram. Their method does not produce a non-negative spectral density estimator, therefore the corresponding estimator of the covariance function is not non-negative definite.

To overcome the drawbacks of previous estimators, in this chapter we propose a new wavelet-based method for the estimation of the spectral density of a Gaussian process and its corresponding covariance function. As a solution to ensure non-negativeness of the spectral density estimator, our method combines the ideas of wavelet thresholding and estimation by information projection. We estimate the spectral density by a projection of the nonlinear wavelet approximation of the periodogram onto a family of exponential functions. Therefore, the estimator is positive by construction. Then, by Bochner's theorem, the corresponding estimator of the covariance function satisfies the non-negative definiteness property. This technique was studied by Barron and Sheu [9] for the approximation of density functions by sequences of exponential families, by Loubes and Yan [60] for penalized maximum likelihood estimation with ℓ_1 penalty, by Antoniadis and Bigot [3] for the study of Poisson inverse problems, and by Bigot and Van Bellegem [15] for log-density deconvolution.

The theoretical optimality of the estimators for the spectral density of a stationary process is generally studied using risk bounds in L_2-norm. This is the case in the papers of Neumann [72], Comte [27] and Fryzlewics, Nason and von Sachs [42] mentioned before. In this work, the behavior of the proposed estimator is established in terms of the rate of convergence of the Kullback-Leibler discrepancy over Besov classes, which is maybe a more natural loss function for the estimation of a spectral density function than the L_2-norm. Moreover, the thresholding rules that we use to derive adaptive estimators differ from previous approaches based on wavelet decomposition and are quite simple to compute. Finally, we compare the performance of our estimator with other estimation procedures on some simulations.

The chapter is organized as follows. Section 2.2 presents the statistical framework under which we work. We define the model, the wavelet-based exponential family and the projection estimators. We also recall the definition of the Kullback-Leibler divergence and some results on Besov spaces. The rate of convergence of the proposed estimators are stated in Section 2.3. Some numerical experiments are described in Section 2.4. Technical lemmas and proofs of the main theorems are gathered at the end of the chapter in Section 2.5.

2.2 Statistical framework

2.2.1 The model

We aim at providing a nonparametric adaptive estimation of the spectral density which satisfies the property of being non-negative in order to guarantee that the covariance estimator is a non-negative definite function. We consider the sequence $(X_t)_{t\in\mathbb{N}}$ that satisfies the following conditions:

Assumption 2.1. *The sequence $(X_1, ..., X_n)$ is an n-sample drawn from a stationary sequence of Gaussian random variables.*

Let σ be the covariance function of the process, i.e. $\sigma(h) = Cov(X_t, X_{t+h})$ with $h \in \mathbb{Z}$. The spectral density f is defined as:

$$f(\omega) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} \sigma(h) e^{-i2\pi\omega h},\ \omega \in [0,1].$$

We need the following standard assumption on σ:

Assumption 2.2. *The covariance function σ is non-negative definite, and there exists two constants $0 < C_1, C_2 < +\infty$ such that $\sum_{h\in\mathbb{Z}} |\sigma(h)| = C_1$ and $\sum_{h\in\mathbb{Z}} |h\sigma^2(h)| = C_2$.*

Assumption 2.2 implies in particular that the spectral density f is bounded by the constant C_1. As a consequence, it is also square integrable. As in Comte [27], the data consist in a number of observations $X_1, ..., X_n$ at regularly spaced points. We want to obtain a positive estimator for the spectral density function f without parametric assumptions on the basis of these observations. For this, we combine the ideas of wavelet thresholding and estimation by information projection.

2.2.2 Estimation by information projection

Wavelet-based exponential family

To ensure non-negativity of the estimator, we will look for approximations over an exponential family. For this, we construct a sieves of exponential functions defined in a wavelet basis.

Let $\phi(\omega)$ and $\psi(\omega)$, respectively, be the scaling and the wavelet functions generated by an orthonormal multiresolution decomposition of $L_2([0,1])$, see Mallat [63] for a detailed exposition on wavelet analysis. Throughout the chapter, the functions ϕ and ψ are supposed to be compactly supported and such that $\|\phi\|_\infty < +\infty$, $\|\psi\|_\infty < +\infty$. Then, for any integer $j_0 \geq 0$, any function $g \in L_2([0,1])$ has the following representation:

$$g(\omega) = \sum_{k=0}^{2^{j_0}-1} \langle g, \phi_{j_0,k} \rangle \phi_{j_0,k}(\omega) + \sum_{j=j_0}^{+\infty} \sum_{k=0}^{2^j-1} \langle g, \psi_{j,k} \rangle \psi_{j,k}(\omega),$$

where $\phi_{j_0,k}(\omega) = 2^{\frac{j_0}{2}} \phi(2^{j_0}\omega - k)$ and $\psi_{j,k}(\omega) = 2^{\frac{j}{2}} \psi(2^j\omega - k)$. The main idea of this chapter is to expand the spectral density f onto this wavelet basis and to find an estimator of this expansion that is then modified to impose the positivity property. The scaling and wavelet coefficients of the spectral density function f are denoted by $a_{j_0,k} = \langle f, \phi_{j_0,k} \rangle$ and $b_{j,k} = \langle f, \psi_{j,k} \rangle$.

To simplify the notations, we write $(\psi_{j,k})_{j=j_0-1}$ for the scaling functions $(\phi_{j,k})_{j=j_0}$. Let $j_1 \geq j_0$ and define the set

$$\Lambda_{j_1} = \left\{ (j,k) : j_0 - 1 \leq j < j_1, 0 \leq k \leq 2^j - 1 \right\}.$$

Note that $|\Lambda_{j_1}| = 2^{j_1}$, where $|\Lambda_{j_1}|$ denotes the cardinality of Λ_{j_1}. Let θ denotes a vector in $\mathbb{R}^{|\Lambda_{j_1}|}$, the wavelet-based exponential family $\mathfrak{E}_{j_1}$ at scale j_1 is defined as the set of functions:

$$\mathfrak{E}_{j_1} = \left\{ f_{j_1,\theta}(.) = \exp\left(\sum_{(j,k)\in\Lambda_{j_1}} \theta_{j,k} \psi_{j,k}(.) \right),\ \theta = (\theta_{j,k})_{(j,k)\in\Lambda_{j_1}} \in \mathbb{R}^{|\Lambda_{j_1}|} \right\}. \tag{2.1}$$

We will enforce our estimator of the spectral density to belong to the family $\mathfrak{E}_{j_1}$ of exponential functions, which are positive by definition.

Information projection

Following Csiszár [29], it is possible to define the projection of a function f onto $\mathfrak{E}_{j_1}$. If this projection exists, it is defined as the function $f_{j_1,\theta^*_{j_1}}$ in the exponential family $\mathfrak{E}_{j_1}$ that is the closest to the true function f in the Kullback-Leibler sense, and is characterized as the unique function in the family $\mathfrak{E}_{j_1}$ for which

$$\left\langle f_{j_1,\theta^*_{j_1}}, \psi_{j,k} \right\rangle = \langle f, \psi_{j,k} \rangle := \beta_{j,k} \text{ for all } (j,k) \in \Lambda_{j_1}.$$

Note that the notation $\beta_{j,k}$ is used to denote both the scaling coefficients $a_{j_0,k}$ and the wavelet coefficients $b_{j,k}$.

Let

$$I_n(\omega) = \frac{1}{2\pi n} \sum_{t=1}^{n} \sum_{t'=1}^{n} \left(X_t - \overline{X} \right) \left(X_{t'} - \overline{X} \right)^* e^{-i2\pi\omega(t-t')},$$

be the classical periodogram, where $\left(X_t - \overline{X} \right)^*$ denotes the conjugate transpose of $\left(X_t - \overline{X} \right)$ and $\overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$. The expansion of $I_n(\omega)$ onto the wavelet basis allows to obtain estimators of $a_{j_0,k}$ and $b_{j,k}$ given by

$$\widehat{a}_{j_0,k} = \int_0^1 I_n(\omega) \phi_{j_0,k}(\omega)\, d\omega \quad \text{and} \quad \widehat{b}_{j,k} = \int_0^1 I_n(\omega) \psi_{j,k}(\omega)\, d\omega. \tag{2.2}$$

It seems therefore natural to estimate the function f by searching for some $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$ such that

$$\left\langle f_{j_1,\widehat{\theta}_n}, \psi_{j,k} \right\rangle = \int_0^1 I_n(\omega) \psi_{j,k}(\omega)\, d\omega := \widehat{\beta}_{j,k} \text{ for all } (j,k) \in \Lambda_{j_1}, \tag{2.3}$$

where $\widehat{\beta}_{j,k}$ denotes both the estimation of the scaling coefficients $\widehat{a}_{j_0,k}$ and the wavelet coefficients $\widehat{b}_{j,k}$. The function $f_{j_1,\widehat{\theta}_n}$ is the spectral density positive projection estimator.

Similarly, the positive nonlinear estimator with hard thresholding is defined as the function $f^{HT}_{j_1,\widehat{\theta}_n,\xi}$ (with $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$) such that

$$\left\langle f^{HT}_{j_1,\widehat{\theta}_n,\xi}, \psi_{j,k} \right\rangle = \delta_\xi \left(\widehat{\beta}_{j,k} \right) \text{ for all } (j,k) \in \Lambda_{j_1}, \tag{2.4}$$

where δ_ξ denotes the hard thresholding rule defined by

$$\delta_\xi(x) = xI(|x| \geq \xi) \text{ for } x \in \mathbb{R},$$

where $\xi > 0$ is an appropriate threshold whose choice is discussed later on.

The existence of these estimators is questionable. Moreover, there is no way to obtain an explicit expression for $\widehat{\theta}_n$. In our simulations, we use a numerical approximation of $\widehat{\theta}_n$ that is obtained via a gradient-descent algorithm with an adaptive step. Proving that such estimators exist with probability one is a difficult task. For the related problem of estimating a density from an independent and identically distributed random variable, it is even shown in Barron and Sheu [9] that for some exponential family (e.g. based on a spline basis), the vector $\widehat{\theta}_n$ may not exist with a small positive probability. Thus, in the next sections, some sufficient conditions are given for the existence of $f_{j_1,\widehat{\theta}_n}$ and $f^{HT}_{j_1,\widehat{\theta}_n,\xi}$ with probability tending to one as $n \to +\infty$.

2.2.3 An appropriate loss function : the Kullback-Leibler divergence

To assess the quality of the estimators, we will measure the discrepancy between an estimator $\widehat{f}$ and the true function f in the sense of relative entropy (Kullback-Leibler divergence) defined by:

$$\Delta\left(f;\widehat{f}\right) = \int_0^1 \left(f \log\left(\frac{f}{\widehat{f}}\right) - f + \widehat{f}\right) d\mu,$$

where μ denotes the Lebesgue measure on $[0,1]$. It can be shown that $\Delta\left(f;\widehat{f}\right)$ is non-negative and equals zero if and only if $\widehat{f} = f$.

2.2.4 Smoothness assumptions

It is well known that Besov spaces for periodic functions in $L_2([0,1])$ can be characterized in terms of wavelet coefficients (see e.g. Mallat [63]). Assume that ψ has m vanishing moments, and let $0 < s < m$ denote the usual smoothness parameter. Then, for a Besov ball $B^s_{p,q}(A)$ of radius $A > 0$ with $1 \leq p,q \leq \infty$, one has that for $s^* = s + 1/2 - 1/p \geq 0$:

$$B^s_{p,q}(A) := \left\{g \in L_2([0,1]) : \|g\|_{s,p,q} \leq A\right\},$$

where

$$\|g\|_{s,p,q} := \left(\sum_{k=0}^{2^{j_0}-1} |\langle g, \phi_{j_0,k}\rangle|^p\right)^{\frac{1}{p}} + \left(\sum_{j=j_0}^{\infty} 2^{js^*q} \left(\sum_{k=0}^{2^j-1} |\langle g, \psi_{j,k}\rangle|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}},$$

with the respective above sums replaced by maximum if $p = \infty$ or $q = \infty$.

The condition that $s + 1/2 - 1/p \geq 0$ is imposed to ensure that $B^s_{p,q}(A)$ is a subspace of $L_2([0,1])$, and we shall restrict ourselves to this case in this chapter (although not always stated, it is clear that all our results hold for $s < m$). Besov spaces allow for more local variability in local smoothness than is typical for functions in the usual Hölder or Sobolev spaces. For instance, a real function f on $[0,1]$ that is piecewise continuous, but for which each piece is locally in C^s, can be an element of $B^s_{p,p}(A)$ with $1 \leq p < 2$, despite the possibility of discontinuities at the transition from one piece to the next (see e.g. Proposition 9.2 in Mallat [63]). Note that if $s > 1$ is not an integer, then $B^s_{2,2}(A)$

is equivalent to a Sobolev ball of order s. Moreover, the space $B^s_{p,q}(A)$ with $1 \leq p < 2$ contains piecewise smooth functions with local irregularities such as discontinuities.

Let $M > 0$ and denote by $F^s_{p,q}(M)$ the set of functions such that

$$F^s_{p,q}(M) = \{f = \exp(g) : \ \|g\|_{s,p,q} \leq M\},$$

where $\|g\|_{s,p,q}$ denotes the norm in the Besov space $B^s_{p,q}$. Assuming that $f \in F^s_{p,q}(M)$ implies that f is strictly positive. In the next section we establish the rate of convergence of our estimators in terms of the Kullback-Leibler discrepancy over Besov classes.

2.3 Asymptotic behavior of the estimators

We make the following assumption on the wavelet basis that guarantees that Assumption 2.2 holds uniformly over $F^s_{p,q}(M)$.

Assumption 2.3. *Let $M > 0$, $1 \leq p \leq 2$ and $s > \frac{1}{p}$. For $f \in F^s_{p,q}(M)$ and $h \in \mathbb{Z}$, let $\sigma(h) = \int\limits_0^1 f(\omega)\, e^{i2\pi\omega h} d\omega$, $C_1(f) = \sum\limits_{h\in\mathbb{Z}} |\sigma(h)|$ and $C_2(f) = \sum\limits_{h\in\mathbb{Z}} |h\sigma^2(h)|$. Then, the wavelet basis is such that there exists a constant $M_* > 0$ such that $C_1(f) \leq M_*$ and $C_2(f) \leq M_*$ for all $f \in F^s_{p,q}(M)$.*

2.3.1 Projection estimation

The following theorem is the general result on the information projection estimator of the spectral density function. Note that the choice of the coarse level resolution level j_0 is of minor importance, and without loss of generality we take $j_0 = 0$ for the estimator $f_{j_1,\widehat{\theta}_n}$.

Theorem 2.1. *Assume that $f \in F^s_{2,2}(M)$ with $s > \frac{1}{2}$ and suppose that Assumptions 2.1, 2.2 and 2.3 are satisfied. Define $j_1 = j_1(n)$ as the largest integer such that $2^{j_1} \leq n^{\frac{1}{2s+1}}$. Then, with probability tending to one as $n \to +\infty$, the information projection estimator (2.3) exists and satisfies:*

$$\Delta\left(f; f_{j_1(n),\widehat{\theta}_n}\right) = \mathcal{O}_p\left(n^{-\frac{2s}{2s+1}}\right).$$

Moreover, the convergence is uniform over the class $F^s_{2,2}(M)$ in the sense that

$$\lim_{K\to+\infty} \lim_{n\to+\infty} \sup_{f\in F^s_{2,2}(M)} \mathbb{P}\left(n^{\frac{2s}{2s+1}} \Delta\left(f; f_{j_1(n),\widehat{\theta}_n}\right) > K\right) = 0.$$

This theorem provides the existence with probability tending to one of a projection estimator for the spectral density f given by $f_{j_1(n),\widehat{\theta}_{j_1(n)}}$. This estimator is strictly positive by construction. Therefore the corresponding estimator of the covariance function $\widehat{\sigma}^L$ (which is obtained as the inverse Fourier transform of $f_{j_1(n),\widehat{\theta}_n}$) is a positive definite function by Bochner's theorem. Hence $\widehat{\sigma}^L$ is a covariance function.

In the related problem of density estimation from an i.i.d. sample, Koo [55] has shown that, for the Kullback-Leibler divergence, $n^{-\frac{2s}{2s+1}}$ is the fastest rate of convergence for the problem of estimating a density f such that $\log(f)$ belongs to the space $B^s_{2,2}(M)$. For spectral densities belonging to a general Besov ball $B^s_{p,q}(M)$, Newman [72] has also shown that $n^{-\frac{2s}{2s+1}}$ is an optimal rate of convergence for the L_2 risk. For the Kullback-Leibler divergence, we conjecture that $n^{-\frac{2s}{2s+1}}$ is the minimax rate of convergence for spectral densities belonging to $F^s_{2,2}(M)$.

The result obtained in the above theorem is nonadaptive because the selection of $j_1(n)$ depends on the unknown smoothness s of f. Moreover, the result is only suited for smooth functions (as $F^s_{2,2}(M)$ corresponds to a Sobolev space of order s) and does not attain an optimal rate of convergence when for example $g = \log(f)$ has singularities. We therefore propose in the next section an adaptive estimator derived by applying an appropriate nonlinear thresholding procedure.

2.3.2 Adaptive estimation

The bound of f is known

In adaptive estimation, we need to define an appropriate thresholding rule for the wavelet coefficients of the periodogram. This threshold is level-dependent and will take the form

$$\xi = \xi_{j,n} = 2\left[2\left\|f\right\|_\infty\left(\sqrt{\frac{\delta\log n}{n}} + 2^{\frac{j}{2}}\left\|\psi\right\|_\infty\frac{\delta\log n}{n}\right) + \frac{C_*}{\sqrt{n}}\right], \tag{2.5}$$

where $\delta \geq 0$ is a tuning parameter whose choice will be discussed later and $C_* = \sqrt{\frac{C_2+39C_1^2}{4\pi^2}}$. The following theorem states that the relative entropy between the true f and its nonlinear estimator achieves in probability the conjectured optimal rate of convergence up to a logarithmic factor over a wide range of Besov balls.

Theorem 2.2. *Assume that $f \in F^s_{p,q}(M)$ with $s > \frac{1}{2} + \frac{1}{p}$ and $1 \leq p \leq 2$. Suppose also that Assumptions 2.1, 2.2 and 2.3 hold. For any $n > 1$, define $j_0 = j_0(n)$ to be the integer such that $2^{j_0} \geq \log n \geq 2^{j_0-1}$, and $j_1 = j_1(n)$ to be the integer such that $2^{j_1} \geq \frac{n}{\log n} \geq 2^{j_1-1}$. For $\delta \geq 6$, take the threshold $\xi_{j,n}$ as in (2.5). Then, the thresholding estimator (2.4) exists with probability tending to one when $n \to +\infty$ and satisfies:*

$$\Delta\left(f; f^{HT}_{j_0(n),j_1(n),\widehat{\theta}_n,\xi_{j,n}}\right) = \mathcal{O}_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).$$

Note that the choices of j_0, j_1 and $\xi_{j,n}$ are independent of the parameter s; hence the estimator $f^{HT}_{j_0(n),j_1(n),\widehat{\theta}_n,\xi_{j,n}}$ is an adaptive estimator which attains in probability what we claim is the optimal rate of convergence, up to a logarithmic factor. In particular, $f^{HT}_{j_0(n),j_1(n),\widehat{\theta}_n,\xi_{j,n}}$ is adaptive on $F^s_{2,2}(M)$. This theorem provides the existence with probability tending to one of a nonlinear estimator for the spectral density. This estimator is strictly positive by construction. Therefore the corresponding estimator of the covariance function $\widehat{\sigma}^{NL}$ (which is obtained as the inverse Fourier transform of $f^{HT}_{j_0(n),j_1(n),\widehat{\theta}_n,\xi_{j,n}}$) is a positive definite function by Bochner's theorem. Hence $\widehat{\sigma}^{NL}$ is a covariance function.

Estimating the bound of f

Although the results of Theorem 2.2 are certainly of some theoretical interest, they are not helpful for practical applications. The (deterministic) threshold $\xi_{j,n}$ depends on the unknown quantities $\|f\|_\infty$ and $C_* := C(C_1, C_2)$, where C_1 and C_2 are unknown constants. To make the method applicable, it is necessary to find some completely data-driven rule for the threshold, which works well over a range as wide as possible of smoothness classes. In this subsection, we give an extension that leads to consider a random threshold which no longer depends on the bound on f neither on C_*. For this, let us consider the dyadic partitions of $[0,1]$ given by $\mathcal{I}_n = \left\{\left(j/2^{J_n}, (j+1)/2^{J_n}\right),\ j = 0, ..., 2^{J_n} - 1\right\}$. Given some positive integer r, we define $\mathcal{P}_n$ as the space of piecewise polynomials of degree r on the dyadic partition $\mathcal{I}_n$ of step 2^{-J_n}. The dimension of $\mathcal{P}_n$ depends on n and is denoted by N_n. Note that $N_n = (r+1)\,2^{J_n}$. This family is regular in the sense that the partition $\mathcal{I}_n$ has equispaced knots.

An estimator of $\|f\|_\infty$ is constructed as proposed by Birgé and Massart [17] in the following way. We take the infinite norm of $\widehat{f_n}$, where $\widehat{f_n}$ denotes the (empirical) orthogonal projection of the periodogram I_n on $\mathcal{P}_n$. We denote by f_n the L_2-orthogonal projection of f on the same space. Then the following theorem holds.

Theorem 2.3. *Assume that $f \in F^s_{p,q}(M)$ with $s > \frac{1}{2} + \frac{1}{p}$ and $1 \le p \le 2$. Suppose also that Assumptions 2.1, 2.2 and 2.3 hold. For any $n > 1$, let $j_0 = j_0(n)$ be the integer such that $2^{j_0} \ge \log n \ge 2^{j_0 - 1}$, and let $j_1 = j_1(n)$ be the integer such that $2^{j_1} \ge \frac{n}{\log n} \ge 2^{j_1 - 1}$. Take the constants $\delta = 6$ and $b \in \left[\frac{3}{4}, 1\right)$, and define the threshold*

$$\widehat{\xi}_{j,n} = 2\left[2\left\|\widehat{f_n}\right\|_\infty \left(\sqrt{\frac{\delta}{(1-b)^2}\frac{\log n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{\delta}{(1-b)^2}\frac{\log n}{n}\right) + \sqrt{\frac{\log n}{n}}\right]. \tag{2.6}$$

Then, if $\|f - f_n\|_\infty \le \frac{1}{4}\|f\|_\infty$ and $N_n \le \frac{\kappa}{(r+1)^2}\frac{n}{\log n}$, where κ is a numerical constant and r is the degree of the polynomials, the thresholding estimator (2.4) exists with probability tending to one as $n \to +\infty$ and satisfies

$$\Delta\left(f; f^{HT}_{j_0(n), j_1(n), \widehat{\theta}_n, \widehat{\xi}_{j,n}}\right) = \mathcal{O}_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).$$

Note that, we finally obtain a fully tractable estimator of f which reaches the optimal rate of convergence without prior knowledge of the regularity of the spectral density, but also which gets rise to a real covariance estimator.

We point out that, in Comte [27] the condition $\|f - f_n\|_\infty \le \frac{1}{4}\|f\|_\infty$ is assumed. Under some regularity conditions on f, results from approximation theory entail that this condition is met. Indeed for $f \in B^s_{p,\infty}$, with $s > \frac{1}{p}$, we know from DeVore and Lorentz [34] that

$$\|f - f_n\|_\infty \le C(s)\,|f|_{s,p}\, N_n^{-\left(s - \frac{1}{p}\right)},$$

with $|f|_{s,p} = \sup\limits_{y>0} y^{-s} w_d(f,y)_p < +\infty$, where $w_d(f,y)_p$ is the modulus of smoothness and

$d = [s] + 1$. Therefore $\|f - f_n\|_\infty \leq \frac{1}{4}\|f\|_\infty$ if $N_n \geq \left(4C(s)\frac{|f|_{s,p}}{\|f\|_\infty}\right)^{\frac{1}{s-\frac{1}{p}}} := C(f,s,p)$, where $C(f,s,p)$ is a constant depending on f, s and p.

2.4 Numerical experiments

In this section we present some numerical experiments which support the claims made in the theoretical part of this chapter. The programs for our simulations were implemented using the MATLAB programming environment. We simulate a time series which is a superposition of an ARMA(2,2) process and a Gaussian white noise:

$$X_t = Y_t + c_0 Z_t, \tag{2.7}$$

where $Y_t + a_1 Y_{t-1} + a_2 Y_{t-2} = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2}$, and $\{\varepsilon_t\}$, $\{Z_t\}$ are independent Gaussian white noise processes with unit variance. The constants were chosen as $a_1 = 0.2$, $a_2 = 0.9$, $b_0 = 1$, $b_1 = 0$, $b_2 = 1$ and $c_0 = 0.5$. We generated a sample of size $n = 1024$ according to (2.7). The spectral density f of (X_t) is shown in Figure 2.1. It has two moderately sharp peaks and is smooth in the rest of the domain.

Starting from the periodogram we considered the Symmlet 8 basis, i.e. the least asymmetric, compactly supported wavelets which are described in Daubechies [31]. We choose j_0 and j_1 as in the hypothesis of Theorem 2.3 and left the coefficients assigned to the father wavelets unthresholded. Hard thresholding is performed using the threshold $\widehat{\xi}_{j,n}$ as in (2.6) for the levels $j = j_0, ..., j_1$, and the empirical coefficients from the higher resolution scales $j > j_1$ are set to zero. This gives the estimate

$$f^{HT}_{j_0,j_1,\xi_{j,n}} = \sum_{k=0}^{2^{j_0}-1} \widehat{a}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} \widehat{b}_{j,k} I\left(\left|\widehat{b}_{j,k}\right| > \xi_{j,n}\right)\psi_{j,k}, \tag{2.8}$$

which is obtained by simply thresholding the wavelet coefficients (2.2) of the periodogram. Note that such an estimator is not guaranteed to be strictly positive in the interval $[0,1]$. However, we use it to built our strictly positive estimator $f^{HT}_{j_0,j_1,\widehat{\theta}_n,\widehat{\xi}_{j,n}}$ (see (2.4) to recall its definition). We want to find $\widehat{\theta}_n$ such that

$$\left\langle f^{HT}_{j_0,j_1,\widehat{\theta}_n,\widehat{\xi}_{j,n}}, \psi_{j,k}\right\rangle = \delta_{\widehat{\xi}_{j,n}}\left(\widehat{\beta}_{j,k}\right) \text{ for all } (j,k) \in \Lambda_{j_1}.$$

For this, we take

$$\widehat{\theta}_n = \underset{\theta \in \mathbb{R}^{\left|\Lambda_{j_1}\right|}}{\arg\min} \sum_{(j,k)\in\Lambda_{j_1}} \left(\langle f_{j_0,j_1,\theta}, \psi_{j,k}\rangle - \delta_{\widehat{\xi}_{j,n}}\left(\widehat{\beta}_{j,k}\right)\right)^2,$$

where $f_{j_0,j_1,\theta}(.) = \exp\left(\sum\limits_{(j,k)\in\Lambda_{j_1}} \theta_{j,k}\psi_{j,k}(.)\right) \in \mathfrak{E}_{j_1}$ and $\mathfrak{E}_{j_1}$ is the family (2.1). To solve this optimization problem we used a gradient descent method with an adaptive step, taking as initial value

$$\theta_0 = \left\langle \log\left(\left(f^{HT}_{j_0,j_1,\widehat{\xi}_{j,n}}\right)_+\right), \psi_{j,k}\right\rangle,$$

where $\left(f^{HT}_{j_0,j_1,\widehat{\xi}_{j,n}}(\omega)\right)_+ := \max\left(f^{HT}_{j_0,j_1,\widehat{\xi}_{j,n}}(\omega), \eta\right)$ for all $\omega \in [0,1]$ and $\eta > 0$ is a small constant.

In Figure 2.1 we display the unconstrained estimator $f^{HT}_{j_0,j_1,\xi_{j,n}}$ as in (2.8), obtained by thresholding of the wavelet coefficients of the periodogram, together with the estimator $f^{HT}_{j_0,j_1,\widehat{\theta}_n,\widehat{\xi}_{j,n}}$, which is strictly positive by construction. Note that these wavelet estimators capture well the peaks and look fairly good on the smooth part too.

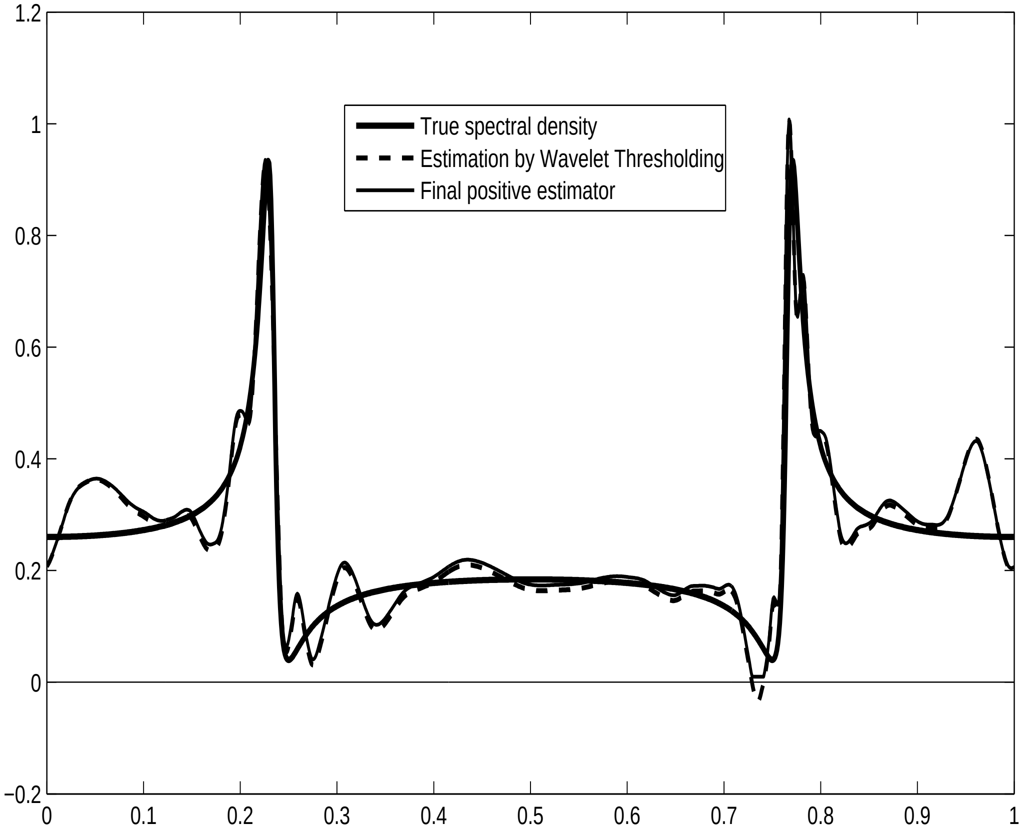


Figure 2.1: True spectral density f, wavelet thresholding estimator $f^{HT}_{j_0,j_1,\widehat{\xi}_{j,n}}$ and final positive estimator $f^{HT}_{j_0,j_1,\widehat{\theta}_n,\widehat{\xi}_{j,n}}$.

We compared our method with the spectral density estimator proposed by Comte [27], which is based on a model selection procedure. As an example, in Comte [27], the author study the behavior of such estimators using a collection of nested models (S_m), with $m = 1, ..., 100$, where S_m is the space of piecewise constant functions, generated by a histogram basis on $[0,1]$ of dimension m with equispaced knots (see Comte [27] for further details). In Figure 2.2 we show the result of this comparison. Note that our method better captures the peaks of the true spectral density.

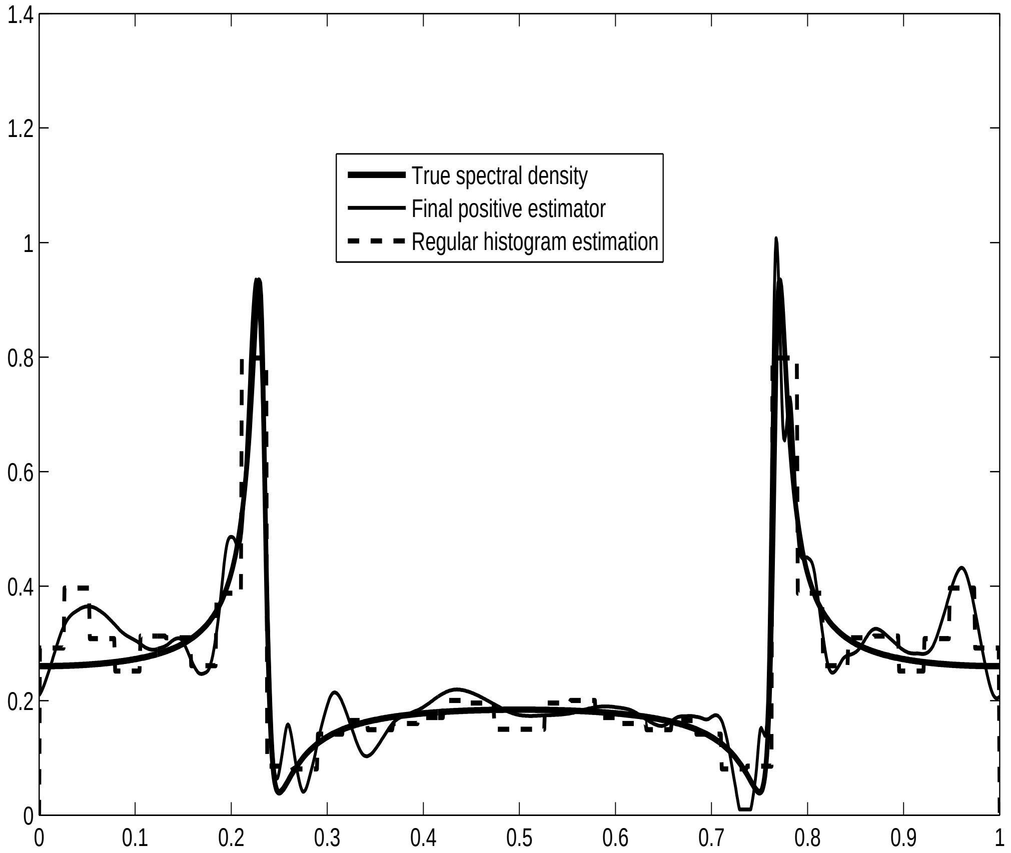


Figure 2.2: True spectral density f, final positive estimator $f^{HT}_{j_0,j_1,\widehat{\theta}_n,\widehat{\xi}_{j,n}}$ and estimator via model selection using regular histograms.

2.5 Proofs

2.5.1 Some notations and definitions.

Throughout all the proofs, C denotes a generic constant whose value may change from line to line. We use the notation $\|.\|_{L_2}$ for the L_2-norm of functions on $[0,1]$. First, let us introduce the following definitions.

Definition 2.1. *Let V_j be the usual multiresolution space at scale j spanned by the scaling functions $(\phi_{j,k})_{0\leq k\leq 2^j-1}$, and define $A_j < +\infty$ as the constant such that $\|\upsilon\|_\infty \leq A_j \|\upsilon\|_{L_2}$ for all $\upsilon \in V_j$.*

Definition 2.2. *For $f \in F^s_{p,q}(M)$, let $g = \log(f)$. Then for all $j \geq j_0 - 1$, define $D_j = \|g - g_j\|_{L_2}$ and $\gamma_j = \|g - g_j\|_\infty$, where $g_j = \sum_{k=0}^{2^j-1} \theta_{j,k}\psi_{j,k}$, with $\theta_{j,k} = \langle g, \psi_{j,k}\rangle$.*

The proof of the following lemma immediately follows from the arguments in the proof of Lemma A.5 in Antoniadis and Bigot [3].

Lemma 2.1. *Let $j \in \mathbb{N}$. Then $A_j \leq C2^{j/2}$. Suppose that $f \in F^s_{p,q}(M)$ with $1 \leq p \leq 2$ and $s > \frac{1}{p}$. Then, uniformly over $F^s_{p,q}(M)$, $D_j \leq C2^{-j(s+1/2-1/p)}$ and $\gamma_j \leq C2^{-j(s-1/p)}$ where C denotes constants depending only on M, s, p and q.*

2.5.2 Technical results on information projection

The estimation of density function based on information projection has been introduced by Barron and Sheu [9]. To apply this method in our context, we recall for completeness a set of results that are useful to prove the existence of our estimators. The proofs of the following lemmas immediately follow from results in Barron and Sheu [9] and Antoniadis and Bigot [3].

Lemma 2.2. *Let f and g two functions in $L_2([0,1])$ such that $\log\left(\frac{f}{g}\right)$ is bounded. Then $\Delta(f;g) \leq \frac{1}{2}e^{\left\|\log\left(\frac{f}{g}\right)\right\|_\infty}\int_0^1 f\left(\log\left(\frac{f}{g}\right)\right)^2 d\mu$, where μ denotes the Lebesgue measure on $[0,1]$.*

Lemma 2.3. *Let $\beta \in \mathbb{R}^{|\Lambda_{j_1}|}$. Assume that there exists some $\theta(\beta) \in \mathbb{R}^{|\Lambda_{j_1}|}$ such that, for all $(j,k) \in \Lambda_{j_1}$, $\theta(\beta)$ is a solution of*

$$\left\langle f_{j,\theta(\beta)}, \psi_{j,k}\right\rangle = \beta_{j,k}.$$

Then for any function f such that $\langle f, \psi_{j,k}\rangle = \beta_{j,k}$ for all $(j,k) \in \Lambda_{j_1}$, and for all $\theta \in \mathbb{R}^{|\Lambda_{j_1}|}$, the following Pythagorian-like identity holds:

$$\Delta(f; f_{j,\theta}) = \Delta\left(f; f_{j,\theta(\beta)}\right) + \Delta\left(f_{j,\theta(\beta)}; f_{j,\theta}\right). \tag{2.9}$$

The next lemma is a key result which gives sufficient conditions for the existence of the vector $\theta(\beta)$ as defined in Lemma 2.3. This lemma also relates distances between the functions in the exponential family to distances between the corresponding wavelet coefficients. Its proof relies upon a series of lemmas on bounds within exponential families for the Kullback-Leibler divergence and can be found in Barron and Sheu [9] and Antoniadis and Bigot [3].

Lemma 2.4. *Let $\theta_0 \in \mathbb{R}^{|\Lambda_{j_1}|}$, $\beta_0 = \left(\beta_{0,(j,k)}\right)_{(j,k)\in\Lambda_{j_1}} \in \mathbb{R}^{|\Lambda_{j_1}|}$ such that $\beta_{0,(j,k)} = \langle f_{j,\theta_0}, \psi_{j,k}\rangle$ for all $(j,k) \in \Lambda_{j_1}$, and $\widetilde{\beta} \in \mathbb{R}^{|\Lambda_{j_1}|}$ a given vector. Let $b = \exp\left(\|\log(f_{j,\theta_0})\|_\infty\right)$ and $e = \exp(1)$. If $\left\|\widetilde{\beta} - \beta_0\right\|_{\ell_2} \leq \frac{1}{2ebA_{j_1}}$ then the solution $\theta\left(\widetilde{\beta}\right)$ of*

$$\langle f_{j_1,\theta}, \psi_{j,k}\rangle = \widetilde{\beta}_{j,k} \text{ for all } (j,k) \in \Lambda_{j_1}$$

exists and satisfies

$$\left\|\theta\left(\widetilde{\beta}\right)-\theta_0\right\|_{\ell_2} \leq 2eb\left\|\widetilde{\beta}-\beta_0\right\|_{\ell_2}$$
$$\left\|\log\left(\frac{f_{j_1,\theta(\beta_0)}}{f_{j_1,\theta(\widetilde{\beta})}}\right)\right\|_{\infty} \leq 2ebA_{j_1}\left\|\widetilde{\beta}-\beta_0\right\|_{\ell_2}$$
$$\Delta\left(f_{j_1,\theta(\beta_0)};f_{j_1,\theta(\widetilde{\beta})}\right) \leq 2eb\left\|\widetilde{\beta}-\beta_0\right\|_{\ell_2}^2,$$

where $\|\beta\|_{\ell_2}$ denotes the standard Euclidean norm for $\beta\in\mathbb{R}^{|\Lambda_{j_1}|}$.

Lemma 2.5. *Suppose that $f\in F^s_{p,q}(M)$ with $s>\frac{1}{p}$ and $1\leq p\leq 2$. Then, there exists a constant M_1 such that for all $f\in F^s_{p,q}(M)$, $0<M_1^{-1}\leq f\leq M_1<+\infty$.*

2.5.3 Technical results for the proofs of the main results

Lemma 2.6. *Let $n\geq 1$, $\beta_{j,k}:=\langle f,\psi_{j,k}\rangle$ and $\widehat{\beta}_{j,k}:=\langle I_n,\psi_{j,k}\rangle$ for $j\geq j_0-1$ and $0\leq k\leq 2^j-1$. Suppose that Assumptions 2.1, 2.2 and 2.3 are satisfied. Then, it holds that $Bias^2\left(\widehat{\beta}_{j,k}\right):=\left(\mathbb{E}\left(\widehat{\beta}_{j,k}\right)-\beta_{j,k}\right)^2\leq\frac{C_*^2}{n}$, where $C_*=\sqrt{\frac{C_2+39C_1^2}{4\pi^2}}$, and it also holds that $\mathbb{V}\left(\widehat{\beta}_{j,k}\right):=\mathbb{E}\left(\widehat{\beta}_{j,k}-\mathbb{E}\left(\widehat{\beta}_{j,k}\right)\right)^2\leq\frac{C}{n}$ for some constant $C>0$. Moreover, there exists a constant $M_2>0$ such that for all $f\in F^s_{p,q}(M)$ with $s>\frac{1}{p}$ and $1\leq p\leq 2$,*

$$\mathbb{E}\left(\widehat{\beta}_{j,k}-\beta_{j,k}\right)^2=Bias^2\left(\widehat{\beta}_{j,k}\right)+\mathbb{V}\left(\widehat{\beta}_{j,k}\right)\leq\frac{M_2}{n}.$$

Proof of Lemma 2.6.

Note that $Bias^2\left(\widehat{\beta}_{j,k}\right)\leq\|f-\mathbb{E}\left(I_n\right)\|^2_{L_2}$. Using Proposition 1 in Comte [27], Assumptions 2.1 and 2.2 imply that $\|f-\mathbb{E}\left(I_n\right)\|^2_{L_2}\leq\frac{C_2+39C_1^2}{4\pi^2 n}$, which gives the result for the bias term. To bound the variance term, remark that

$$\begin{aligned}\mathbb{V}\left(\widehat{\beta}_{j,k}\right)&=\mathbb{E}\left(\langle I_n-\mathbb{E}\left(I_n\right),\psi_{j,k}\rangle\right)^2\\&\leq\mathbb{E}\left(\|I_n-\mathbb{E}\left(I_n\right)\|^2_{L_2}\|\psi_{j,k}\|^2_{L_2}\right)\\&=\int_0^1\mathbb{E}|I_n(\omega)-\mathbb{E}\left(I_n(\omega)\right)|^2d\omega.\end{aligned}$$

Then, under Assumptions 2.1 and 2.2, it follows that there exists an absolute constant $C>0$ such that for all $\omega\in[0,1]$, $\mathbb{E}|I_n(\omega)-\mathbb{E}\left(I_n(\omega)\right)|^2\leq\frac{C}{n}$. To complete the proof it remains to remark that Assumption 2.3 implies that these bounds for the bias and the variance hold uniformly over $F^s_{p,q}(M)$. □

Lemma 2.7. *Let $n\geq 1$, $b_{j,k}:=\langle f,\psi_{j,k}\rangle$ and $\widehat{b}_{j,k}:=\langle I_n,\psi_{j,k}\rangle$ for $j\geq j_0$ and $0\leq k\leq 2^j-1$. Suppose that Assumptions 2.1 and 2.2 hold. Then for any $x>0$,*

$$\mathbb{P}\left(|\widehat{b}_{j,k}-b_{j,k}|>2\|f\|_\infty\left(\sqrt{\frac{x}{n}}+2^{j/2}\|\psi\|_\infty\frac{x}{n}\right)+\frac{C_*}{\sqrt{n}}\right)\leq 2e^{-x},$$

where $C_*=\sqrt{\frac{C_2+39C_1^2}{4\pi^2}}$.

Proof of Lemma 2.7.

Note that

$$\widehat{b}_{j,k} = \frac{1}{2\pi n}\sum_{t=1}^{n}\sum_{t'=1}^{n}\left(X_t-\overline{X}\right)\left(X_{t'}-\overline{X}\right)^{*}\int_{0}^{1} e^{-i2\pi\omega(t-t')}\psi_{j,k}\left(\omega\right)d\omega = \frac{1}{2\pi n}X^{\top}\mathbf{T}_n(\psi_{j,k})X^{*},$$

where $X = \left(X_1-\overline{X},...,X_n-\overline{X}\right)^{\top}$, $X^{\top}$ denotes the transpose of X and $\mathbf{T}_n(\psi_{j,k})$ is the Toeplitz matrix with entries $[\mathbf{T}_n(\psi_{j,k})]_{t,t'} = \int_0^1 e^{-i2\pi\omega(t-t')}\psi_{j,k}\left(\omega\right)d\omega$, $1 \le t,t' \le n$. We can assume without loss of generality that $\mathbb{E}\left(X_t\right) = 0$, and then under Assumptions 2.1 and 2.2, X is a centered Gaussian vector in $\mathbb{R}^n$ with covariance matrix $\mathbf{\Sigma} = \mathbf{T}_n(f)$. Using the decomposition $X = \mathbf{\Sigma}^{\frac{1}{2}}\varepsilon$, where $\varepsilon \sim N\left(0,\mathbf{I}_n\right)$, it follows that $\widehat{b}_{j,k} = \frac{1}{2\pi n}\varepsilon^{\top}\mathbf{A}_{j,k}\varepsilon$, with $\mathbf{A}_{j,k} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{T}_n(\psi_{j,k})\mathbf{\Sigma}^{\frac{1}{2}}$. Note also that $\mathbb{E}\left(\widehat{b}_{j,k}\right) = \frac{1}{2\pi n}\mathrm{Tr}\left(\mathbf{A}_{j,k}\right)$, where $\mathrm{Tr}\left(\mathbf{A}\right)$ denotes the trace of a matrix $\mathbf{A}$.

Now let $s_1,\ldots,s_n$ be the eigenvalues of the Hermitian matrix $\mathbf{A}_{j,k}$ ordered by their absolute values, that is $|s_1| \ge |s_2| \ge \ldots \ge |s_n|$, and let

$$Z = 2\pi n\left(\widehat{b}_{j,k} - \mathbb{E}\left(\widehat{b}_{j,k}\right)\right) = \varepsilon^{\top}\mathbf{A}_{j,k}\varepsilon - \mathrm{Tr}\left(\mathbf{A}_{j,k}\right).$$

Then, for $0 < \lambda < (2|s_1|)^{-1}$ one has that

$$\begin{aligned}\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right) &= \sum_{i=1}^{n} -\lambda s_i - \frac{1}{2}\log\left(1-2\lambda s_i\right)\\ &= \sum_{i=1}^{n}\sum_{\ell=2}^{+\infty}\frac{1}{2\ell}(2s_i\lambda)^{\ell} \le \sum_{i=1}^{n}\sum_{\ell=2}^{+\infty}\frac{1}{2\ell}(2|s_i|\lambda)^{\ell}\\ &\le \sum_{i=1}^{n} -\lambda|s_i| - \frac{1}{2}\log\left(1-2\lambda|s_i|\right),\end{aligned}$$

where we have used the fact that $-\log(1-x) = \sum_{\ell=1}^{+\infty}\frac{x^{\ell}}{\ell}$ for $x < 1$. Then using the inequality $-u - \frac{1}{2}\log\left(1-2u\right) \le \frac{u^2}{1-2u}$ that holds for all $0 < u < \frac{1}{2}$, the above inequality implies that

$$\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right) \le \sum_{i=1}^{n}\frac{\lambda^2|s_i|^2}{1-2\lambda|s_i|} \le \frac{\lambda^2\|s\|_{\ell_2}^2}{1-2\lambda|s_1|},$$

where $\|s\|_{\ell_2}^2 = \sum_{i=1}^{n}|s_i|^2$. Arguing as in Birgé and Massart [17], the above inequality implies that for any $x > 0$, $\mathbb{P}(|Z| > 2|s_1|x + 2\|s\|_{\ell_2}\sqrt{x}) \le 2e^{-x}$, which can also be written as

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - \mathbb{E}\left(\widehat{b}_{j,k}\right)\right| > 2|s_1|\frac{x}{2\pi n} + 2\frac{\|s\|_{\ell_2}}{2\pi n}\sqrt{x}\right) \le 2e^{-x},$$

that implies

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - \mathbb{E}\left(\widehat{b}_{j,k}\right)\right| > 2|s_1|\frac{x}{n} + 2\frac{\|s\|_{\ell_2}}{n}\sqrt{x}\right) \le 2e^{-x}. \tag{2.10}$$

Let $\tau\left(\mathbf{A}\right)$ denotes the spectral radius of a square matrix $\mathbf{A}$. For the Toeplitz matrices $\mathbf{\Sigma} = \mathbf{T}_n(f)$ and $\mathbf{T}_n(\psi_{j,k})$ one has that

$$\tau\left(\mathbf{\Sigma}\right) \leq \|f\|_\infty \text{ and } \tau\left(\mathbf{T}_n(\psi_{j,k})\right) \leq \|\psi_{j,k}\|_\infty = 2^{j/2}\|\psi\|_\infty.$$

The above inequalities imply that

$$|s_1| = \tau\left(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{T}_n(\psi_{j,k})\mathbf{\Sigma}^{\frac{1}{2}}\right) \leq \tau\left(\mathbf{\Sigma}\right)\tau\left(\mathbf{T}_n(\psi_{j,k})\right) \leq \|f\|_\infty 2^{j/2}\|\psi\|_\infty. \tag{2.11}$$

Let λ_i, $i = 1.,,,.n$, be the eigenvalues of $\mathbf{T}_n(\psi_{j,k})$. From Lemma 3.1 in Davies [33], we have that

$$\limsup_{n\to+\infty}\frac{1}{n}\mathrm{Tr}\left(\mathbf{T}_n(\psi_{j,k})^2\right) = \limsup_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\lambda_i^2 = \int\limits_0^1 \psi_{j,k}^2\left(\omega\right)d\omega = 1,$$

which implies that

$$\begin{aligned} \|s\|_{\ell_2}^2 &= \sum_{i=1}^{n}|s_i|^2 = \mathrm{Tr}\left(\mathbf{A}_{j,k}^2\right) = \mathrm{Tr}\left(\left(\mathbf{\Sigma}\mathbf{T}_n(\psi_{j,k})\right)^2\right) \\ &\leq \tau\left(\mathbf{\Sigma}\right)^2\mathrm{Tr}\left(\mathbf{T}_n(\psi_{j,k})^2\right) \leq \|f\|_\infty^2 n, \end{aligned} \tag{2.12}$$

where we have used the inequality $\mathrm{Tr}\left((\mathbf{AB})^2\right) \leq \tau\left(\mathbf{A}\right)^2\mathrm{Tr}\left(\mathbf{B}^2\right)$ that holds for any pair of Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$. Combining (2.10), (2.11) and (2.12), we finally obtain that for any $x > 0$

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - \mathbb{E}\left(\widehat{b}_{j,k}\right)\right| > 2\|f\|_\infty\left(\sqrt{\frac{x}{n}} + 2^{j/2}\|\psi\|_\infty\frac{x}{n}\right)\right) \leq 2e^{-x}. \tag{2.13}$$

Now, let $\xi_{j,n} = 2\|f\|_\infty\left(\sqrt{\frac{x}{n}} + 2^{j/2}\|\psi\|_\infty\frac{x}{n}\right) + \frac{C_*}{\sqrt{n}}$, and note that

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \xi_{j,n}\right) \leq \mathbb{P}\left(\left|\widehat{b}_{j,k} - \mathbb{E}\left(\widehat{b}_{j,k}\right)\right| > \xi_{j,n} - \left|\mathbb{E}\left(\widehat{b}_{j,k}\right) - b_{j,k}\right|\right).$$

By Lemma 2.6, one has that $\left|\mathbb{E}\left(\widehat{b}_{j,k}\right) - b_{j,k}\right| \leq \frac{C_*}{\sqrt{n}}$. Thus, $\xi_{j,n} - \left|\mathbb{E}\left(\widehat{b}_{j,k}\right) - b_{j,k}\right| \geq \xi_{j,n} - \frac{C_*}{\sqrt{n}}$, which implies using (2.13) that

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \xi_{j,n}\right) \leq \mathbb{P}\left(\left|\widehat{b}_{j,k} - \mathbb{E}\left(\widehat{b}_{j,k}\right)\right| > \xi_{j,n} - \frac{C_*}{\sqrt{n}}\right) \leq 2e^{-x},$$

which completes the proof of Lemma 2.7. □

Lemma 2.8. *Assume that $f \in F_{p,q}^s\left(M\right)$ with $s > \frac{1}{2} + \frac{1}{p}$ and $1 \leq p \leq 2$. Suppose that Assumptions 2.1, 2.2 and 2.3 hold. For any $n > 1$, define $j_0 = j_0\left(n\right)$ to be the integer such that $2^{j_0} > \log n \geq 2^{j_0-1}$, and $j_1 = j_1\left(n\right)$ to be the integer such that $2^{j_1} \geq \frac{n}{\log n} \geq 2^{j_1-1}$. For $\delta \geq 6$, take the threshold $\xi_{j,n} = 2\left[2\left\|f\right\|_\infty\left(\sqrt{\frac{\delta\log n}{n}} + 2^{\frac{j}{2}}\left\|\psi\right\|_\infty\frac{\delta\log n}{n}\right) + \frac{C_*}{\sqrt{n}}\right]$ as in (2.5),* *where* $C_* = \sqrt{\frac{C_2+39C_1^2}{4\pi^2}}$. *Let* $\beta_{j,k} := \langle f, \psi_{j,k}\rangle$ *and* $\widehat{\beta}_{\xi_{j,n},(j,k)} := \delta_{\xi_{j,n}}\left(\widehat{\beta}_{j,k}\right)$ *with* $(j,k) \in \Lambda_{j_1}$ *as in* (2.4). *Take* $\beta = (\beta_{j,k})_{(j,k)\in\Lambda_{j_1}}$ *and* $\widehat{\beta}_{\xi_{j,n}} = \left(\widehat{\beta}_{\xi_{j,n},(j,k)}\right)_{(j,k)\in\Lambda_{j_1}}$. *Then there exists a constant* $M_3 > 0$ *such that for all sufficiently large* n*:*

$$\mathbb{E}\left\|\beta - \widehat{\beta}_{\xi_{j,n}}\right\|_{\ell_2}^2 := \mathbb{E}\left(\sum_{(j,k)\in\Lambda_{j_1}}\left|\beta_{j,k} - \delta_{\xi_{j,n}}\left(\widehat{\beta}_{j,k}\right)\right|^2\right) \leq M_3\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}$$

uniformly over $F^s_{p,q}(M)$.

Proof of Lemma 2.8.

Taking into account that

$$\begin{aligned}\mathbb{E}\left\|\beta - \widehat{\beta}_{\xi_{j,n}}\right\|_{\ell_2}^2 &= \sum_{k=0}^{2^{j_0}-1}\mathbb{E}\left(a_{j_0,k} - \widehat{a}_{j_0,k}\right)^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^2 I\left(\left|\widehat{b}_{j,k}\right| > \xi_{j,n}\right)\right] \\ &+ \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} b_{j,k}^2\mathbb{P}\left(\left|\widehat{b}_{j,k}\right| \leq \xi_{j,n}\right) \\ &:= T_1 + T_2 + T_3,\end{aligned} \tag{2.14}$$

we are interested in bounding these three terms. The bound for T_1 follows from Lemma 2.6 and the fact that $j_0 = \log_2(\log n) \leq \frac{1}{2s+1}\log_2(n)$:

$$T_1 = \sum_{k=0}^{2^{j_0}-1}\mathbb{E}\left(a_{j_0,k} - \widehat{a}_{j_0,k}\right)^2 = O\left(\frac{2^{j_0}}{n}\right) \leq O\left(n^{-\frac{2s}{2s+1}}\right). \tag{2.15}$$

To bound T_2 and T_3 we proceed as in Härdle, Kerkyacharian, Picard and Tsybakov [48]. Write

$$T_2 = \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^2\left\{I\left(\left|\widehat{b}_{j,k}\right| > \xi_{j,n}\,, |b_{j,k}| > \frac{\xi_{j,n}}{2}\right) + I\left(\left|\widehat{b}_{j,k}\right| > \xi_{j,n}\,, |b_{j,k}| \leq \frac{\xi_{j,n}}{2}\right)\right\}\right]$$

and

$$T_3 = \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} b_{j,k}^2\left[\mathbb{P}\left(\left|\widehat{b}_{j,k}\right| \leq \xi_{j,n}\,, |b_{j,k}| \leq 2\xi_{j,n}\right) + \mathbb{P}\left(\left|\widehat{b}_{j,k}\right| \leq \xi_{j,n}, |b_{j,k}| > 2\xi_{j,n}\right)\right].$$

Note that

$$\begin{aligned} I\left(\left|\widehat{b}_{j,k}\right| > \xi_{j,n}\,, |b_{j,k}| \leq \frac{\xi_{j,n}}{2}\right) &\leq I\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right), \\ I\left(\left|\widehat{b}_{j,k}\right| \leq \xi_{j,n}\,, |b_{j,k}| > 2\xi_{j,n}\right) &\leq I\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right),\end{aligned} \tag{2.16}$$

and if $\left|\widehat{b}_{j,k}\right| \leq \xi_{j,n}$, $|b_{j,k}| > 2\xi_{j,n}$, then $\left|\widehat{b}_{j,k}\right| \leq \frac{|b_{j,k}|}{2}$, and $\left|\widehat{b}_{j,k} - b_{j,k}\right| \geq \left|\widehat{b}_{j,k}\right| - |b_{j,k}| \geq \frac{|b_{j,k}|}{2}$. Therefore

$$b_{j,k}^2 \leq 4\left(\widehat{b}_{j,k} - b_{j,k}\right)^2. \tag{2.17}$$

Using (2.16) and (2.17), we get

$$\begin{aligned}T_2 + T_3 &\leq \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left\{\left(b_{j,k} - \widehat{b}_{j,k}\right)^2\right\} I\left(|b_{j,k}| > \frac{\xi_{j,n}}{2}\right) + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} b_{j,k}^2 I\left(|b_{j,k}| \leq 2\xi_{j,n}\right)\\ &+ 5\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left\{\left(b_{j,k} - \widehat{b}_{j,k}\right)^2 I\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right)\right\}\\ &:= T' + T'' + T'''.\end{aligned}$$

Now we bound T'''. Using Cauchy-Schwarz inequality, we obtain

$$T''' \leq 5\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}^{\frac{1}{2}}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^4\right]\mathbb{P}^{\frac{1}{2}}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right).$$

By the same inequality we get

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{b}_{j,k} - b_{j,k}\right)^4\right] &= \mathbb{E}\left[\langle I_n - f, \psi_{j,k}\rangle^4\right]\\ &\leq \mathbb{E}\left[\|I_n - f\|_{L_2}^4 \|\psi_{j,k}\|_{L_2}^4\right]\\ &= O\left(\mathbb{E}\|I_n - f\|_{L_2}^4\right).\end{aligned}$$

It can be checked that $\mathbb{E}\|I_n - f\|_{L_2}^4 \leq 8\mathbb{E}\left(\|I_n - \mathbb{E}(I_n)\|_{L_2}^4 + \|\mathbb{E}(I_n) - f\|_{L_2}^4\right)$. According to Comte [27], $\mathbb{E}\|I_n - \mathbb{E}(I_n)\|_{L_2}^4 = O(n^2)$. From the proof of Lemma 2.6 we get that $\|\mathbb{E}(I_n) - f\|_{L_2}^4 = O\left(\frac{1}{n^2}\right)$. Therefore $\mathbb{E}\|I_n - f\|_{L_2}^4 \leq O\left(n^2 + \frac{1}{n^2}\right) = O(n^2)$. Hence

$$\mathbb{E}\left[\left(\widehat{b}_{j,k} - b_{j,k}\right)^4\right] = O\left(\mathbb{E}\|I_n - f\|_{L_2}^4\right) = O\left(n^2\right).$$

For the bound of $\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right)$ we use the result of Lemma 2.7 with $x = \delta\log n$, where $\delta > 0$ is a constant to be specified later. We obtain

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right) &= \mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > 2\|f\|_\infty\left(\sqrt{\frac{\delta\log n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty\frac{\delta\log n}{n}\right) + \frac{C_*}{\sqrt{n}}\right)\\ &\leq 2e^{-\delta\log n} = 2n^{-\delta}.\end{aligned}$$

Therefore, for $\delta \geq 6$, we get

$$\begin{aligned}T''' &\leq 5\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}^{\frac{1}{2}}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^4\right]\mathbb{P}^{\frac{1}{2}}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\xi_{j,n}}{2}\right)\\ &\leq C\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} n^{1-\frac{\delta}{2}} \leq Cn^{-2}\sum_{j=j_0}^{j_1}2^j = O\left(n^{-2}2^{j_1}\right) = O\left(\frac{n^{-1}}{\log n}\right) \leq O\left(n^{-\frac{2s}{2s+1}}\right).\end{aligned}$$

Now we follow results found in Pensky and Sapatinas [75] to bound T' and T''. Let j_A be the integer such that $2^{j_A} > \left(\frac{n}{\log n}\right)^{\frac{1}{2s+1}} > 2^{j_A-1}$ (note that given our assumptions $j_0 \leq j_A \leq j_1$ for all sufficiently large n), then T' can be partitioned as $T' = T'_1 + T'_2$, where the first component is calculated over the set of indices $j_0 \leq j \leq j_A$ and the second component over $j_A + 1 \leq j \leq j_1$. Hence, using Lemma 2.6 we obtain

$$T'_1 \leq C \sum_{j=j_0}^{j_A} \frac{2^j}{n} = O\left(2^{j_A} n^{-1}\right) = O\left(\left(\frac{n}{\log n}\right)^{\frac{1}{2s+1}} n^{-1}\right) \leq O\left(n^{-\frac{2s}{2s+1}}\right). \tag{2.18}$$

To obtain a bound for T'_2, we will use that if $f \in F^s_{p,q}(M)$, then for some constant C, dependent on s, p, q and $M > 0$ only, we have that

$$\sum_{k=0}^{2^j-1} b^2_{j,k} \leq C 2^{-2js^*}, \tag{2.19}$$

for $1 \leq p \leq 2$, where $s^* = s + \frac{1}{2} - \frac{1}{p}$. Taking into account that $I\left(|b_{j,k}| > \frac{\xi_{j,n}}{2}\right) \leq \frac{4}{\xi^2_{j,n}} |b_{j,k}|^2$, we get

$$\begin{aligned}
T'_2 &\leq \frac{C}{n} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{4}{\xi^2_{j,n}} |b_{j,k}|^2 \leq \frac{C}{n} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{|b_{j,k}|^2}{\left[2 \|f\|_\infty \left(\sqrt{\frac{\delta \log n}{n}} + 2^{\frac{j}{2}} \|\psi\|_\infty \frac{\delta \log n}{n}\right) + \frac{C_*}{\sqrt{n}}\right]^2} \\
&\leq \frac{C}{n} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{|b_{j,k}|^2}{4 \|f\|^2_\infty \left(\sqrt{\frac{\delta \log n}{n}} + 2^{\frac{j_A}{2}} \|\psi\|_\infty \frac{\delta \log n}{n}\right)^2} \\
&\leq \frac{C\left(\|f\|_\infty\right)}{\left(\sqrt{\delta \log n} + \|\psi\|_\infty \delta n^{\frac{-s}{2s+1}} (\log n)^{\frac{4s+1}{4s+2}}\right)^2} 2^{-2s^* j_A} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} 2^{2js^*} |b_{j,k}|^2 \\
&\leq O\left(2^{-2s^* j_A}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s^*}{2s+1}}\right),
\end{aligned}$$

where we used the fact that $\sqrt{\delta \log n} + \|\psi\|_\infty \delta n^{\frac{-s}{2s+1}} (\log n)^{\frac{4s+1}{4s+2}} \to +\infty$ when $n \to +\infty$. Now remark that if $p = 2$ then $s^* = s$ and thus

$$T'_2 = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s^*}{2s+1}}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right). \tag{2.20}$$

For the case $1 \leq p < 2$, the repeated use of the fact that if $B, D > 0$ then $I\left(|b_{j,k}| > B + D\right) \leq$

$I\left(|b_{j,k}| > B\right)$, allow us to obtain that

$$\begin{aligned}
T_2' &\leq C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{1}{n} I\left(|b_{j,k}| > \frac{\xi_{j,n}}{2}\right) \leq C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{1}{n} I\left(|b_{j,k}| > 2\left\|f\right\|_{\infty} \sqrt{\frac{\delta \log n}{n}}\right) \\
&= C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{1}{n} \left|b_{j,k}\right|^{-p} \left|b_{j,k}\right|^{p} I\left(\left|b_{j,k}\right|^{-p} < \left(2\left\|f\right\|_{\infty} \sqrt{\delta}\sqrt{\frac{\log n}{n}}\right)^{-p}\right) \\
&\leq C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{1}{n} \left(2\left\|f\right\|_{\infty} \sqrt{\delta}\sqrt{\frac{\log n}{n}}\right)^{-p} \left|b_{j,k}\right|^{p}.
\end{aligned}$$

Since $f \in F_{p,q}^{s}(M)$ it follows that there exists a constant C depending only on p, q, s and A such that

$$\sum_{k=0}^{2^j-1} \left|b_{j,k}\right|^{p} \leq C 2^{-pjs^*}, \tag{2.21}$$

where $s^* = s + \frac{1}{2} - \frac{1}{p}$ as before. By (2.21) we get

$$\begin{aligned}
T_2' &\leq (\log n)\, C\left(\left\|f\right\|_{\infty}, \delta, p\right) \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} \frac{(\log n)^{-\frac{p}{2}}}{n^{1-\frac{p}{2}}} \left|b_{j,k}\right|^{p} \\
&\leq C\left(\left\|f\right\|_{\infty}, \delta, p\right) \frac{(\log n)^{1-\frac{p}{2}}}{n^{1-\frac{p}{2}}} \sum_{j=j_A}^{j_1} C 2^{-pjs^*} = O\left(\frac{(\log n)^{1-\frac{p}{2}}}{n^{1-\frac{p}{2}}} 2^{-pj_A s^*}\right) \\
&= O\left(\left(\frac{n}{\log n}\right)^{\frac{p}{2}-1} \left(\frac{n}{\log n}\right)^{-\frac{p\left(s+\frac{1}{2}-\frac{1}{p}\right)}{2s+1}}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).
\end{aligned} \tag{2.22}$$

Hence, by (2.18), (2.20) and (2.22), $T' = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right)$.

Now, set j_A as before, then T'' can be split into $T'' = T_1'' + T_2''$, where the first component is calculated over the set of indices $j_0 \leq j \leq j_A$ and the second component over $j_A + 1 \leq j \leq j_1$. Then

$$\begin{aligned}
T_1'' &\leq \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} b_{j,k}^2 I\left(\left|b_{j,k}\right|^2 \leq 32\left[4\left\|f\right\|_{\infty}^2 \left(\sqrt{\frac{\delta \log n}{n}} + 2^{\frac{j}{2}} \left\|\psi\right\|_{\infty} \frac{\delta \log n}{n}\right)^2 + \frac{C_*^2}{n}\right]\right) \\
&\leq C\left(\left\|f\right\|_{\infty}\right) \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \left(\sqrt{\frac{\delta \log n}{n}} + 2^{\frac{j}{2}} \left\|\psi\right\|_{\infty} \frac{\delta \log n}{n}\right)^2 + C\left(C_*\right) \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \frac{1}{n},
\end{aligned}$$

where we used that $(B+D)^2 \leq 2\left(B^2 + D^2\right)$ for all $B, D \in \mathbb{R}$. Using the same property again, we obtain the desired bound for T_1'':

$$\begin{aligned}
T_1'' &\leq C\left(\|f\|_\infty\right) \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \left(\frac{\delta \log n}{n} + 2^j \|\psi\|_\infty^2 \frac{\delta^2 (\log n)^2}{n^2} \right) + C\left(C_*\right) \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \frac{1}{n} \\
&\leq C\left(\|f\|_\infty, \delta\right) \frac{\log n}{n} \sum_{j=j_0}^{j_A} 2^j + C\left(\|f\|_\infty, \delta, \|\psi\|_\infty\right) \frac{(\log n)^2}{n^2} \sum_{j=j_0}^{j_A} 2^{2j} + \frac{C\left(C_*\right)}{n} \sum_{j=j_0}^{j_A} 2^j \\
&\leq C\left(\|f\|_\infty, \delta, C_*\right) \frac{\log n}{n} 2^{j_A} + C\left(\|f\|_\infty, \delta, \|\psi\|_\infty\right) \frac{(\log n)^2}{n^2} 2^{2j_A} \\
&= O\left((\log n)^{\frac{2s}{2s+1}} n^{-\frac{2s}{2s+1}} + (\log n)^{\frac{4s}{2s+1}} n^{-\frac{4s}{2s+1}} \right) \leq O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s}{2s+1}} \right). \qquad (2.23)
\end{aligned}$$

To bound T_2'', we proceed as follows:

$$T_2'' \leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} b_{j,k}^2 = O\left(2^{-2j_A s^*}\right) = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s^*}{2s+1}} \right),$$

where we have used the condition (2.19). Now remark that if $p = 2$ then $s^* = s$ and thus

$$T_2'' = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s^*}{2s+1}} \right) = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s}{2s+1}} \right). \qquad (2.24)$$

If $1 \leq p < 2$,

$$\begin{aligned}
T_2'' &= \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^{2-p} |b_{j,k}|^p I\left(|b_{j,k}| \leq 2\xi_{j,n}\right) \\
&\leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} \left(8\|f\|_\infty \sqrt{\frac{\delta \log n}{n}} + 8\|f\|_\infty 2^{\frac{j_1}{2}} \|\psi\|_\infty \frac{\delta \log n}{n} + 4\sqrt{\frac{\log n}{n}} \right)^{2-p} |b_{j,k}|^p \\
&= \left(C\left(\|f\|_\infty, \delta\right) + 4 + C\left(\|f\|_\infty, \|\psi\|_\infty, \delta\right)\right)^{2-p} \left(\sqrt{\frac{\log n}{n}} \right)^{2-p} \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^p \\
&= O\left(\left(\frac{\log n}{n} \right)^{\frac{2-p}{2}} 2^{-pj_A s^*} \right) = O\left(\left(\frac{n}{\log n} \right)^{\frac{p}{2}-1-\frac{p\left(s+\frac{1}{2}-\frac{1}{p}\right)}{2s+1}} \right) = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s}{2s+1}} \right),
\end{aligned} \qquad (2.25)$$

where we have used condition (2.21) and the fact that $C_* \leq \sqrt{\log n}$ for n sufficiently large, taking into account that the constant $C_* := C\left(C_1, C_2\right)$ does not depend on n. Hence, by (2.23), (2.24) and (2.25), $T'' = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s}{2s+1}} \right)$. Combining all terms in (2.14), we conclude that:

$$\mathbb{E} \left\| \beta - \widehat{\beta}_{\xi_{j,n}} \right\|_{\ell_2}^2 = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s}{2s+1}} \right).$$

This completes the proof. □

Lemma 2.9. *Assume that $f \in F^s_{p,q}(M)$ with $s > \frac{1}{2} + \frac{1}{p}$ and $1 \leq p \leq 2$. Suppose that Assumptions 2.1, 2.2 and 2.3 hold. For any $n > 1$, define $j_0 = j_0(n)$ to be the integer such that $2^{j_0} > \log n \geq 2^{j_0 - 1}$, and $j_1 = j_1(n)$ to be the integer such that $2^{j_1} \geq \frac{n}{\log n} \geq 2^{j_1 - 1}$. Define the threshold*

$$\widehat{\xi}_{j,n} = 2\left[2\left\|\widehat{f}_n\right\|_\infty\left(\sqrt{\frac{\delta}{(1-b)^2}\frac{\log n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty\frac{\delta}{(1-b)^2}\frac{\log n}{n}\right) + \sqrt{\frac{\log n}{n}}\right]$$

as in (2.6) with $\delta = 6$ and $b \in \left[\frac{3}{4}, 1\right)$. Let $\beta_{j,k} := \langle f, \psi_{j,k}\rangle$ and $\widehat{\beta}_{\widehat{\xi}_{j,n},(j,k)} := \delta_{\widehat{\xi}_{j,n}}\left(\widehat{\beta}_{j,k}\right)$ with $(j,k) \in \Lambda_{j_1}$ as in (2.4). Take $\beta = (\beta_{j,k})_{(j,k)\in\Lambda_{j_1}}$ and $\widehat{\beta}_{\widehat{\xi}_{j,n}} = \left(\widehat{\beta}_{\widehat{\xi}_{j,n},(j,k)}\right)_{(j,k)\in\Lambda_{j_1}}$. Then, if $\|f - f_n\|_\infty \leq \frac{1}{4}\|f\|_\infty$ and $N_n \leq \frac{\kappa}{(r+1)^2}\frac{n}{\log n}$, where κ is a numerical constant and r is the degree of the polynomials, there exists a constant $M_4 > 0$ such that for all sufficiently large n:

$$\mathbb{E}\left\|\beta - \widehat{\beta}_{\widehat{\xi}_{j,n}}\right\|^2_{\ell_2} := \mathbb{E}\left(\sum_{(j,k)\in\Lambda_{j_1}}\left|\delta_{\widehat{\xi}_{j,n}}\left(\widehat{\beta}_{j,k}\right) - \beta_{j,k}\right|^2\right) \leq M_4\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}$$

uniformly over $F^s_{p,q}(M)$.

Proof of Lemma 2.9.

Recall that f_n is the L_2 orthogonal projection of f on the space $\mathcal{P}_n$ of piecewise polynomials of degree r on a dyadic partition with step 2^{-J_n}. The dimension of $\mathcal{P}_n$ is $N_n = (r+1)2^{J_n}$. Let $\widehat{f_n}$ similarly be the orthogonal projection of I_n on $\mathcal{P}_n$. By doing analogous work as the one done to obtain (2.14), we get that

$$\mathbb{E}\left\|\beta - \widehat{\beta}_{\widehat{\xi}_{j,n}}\right\|^2_{\ell_2} := T_1 + T_2 + T_3, \tag{2.26}$$

where $T_1 = \sum\limits_{k=0}^{2^{j_0}-1}\mathbb{E}\left(a_{j_0,k} - \widehat{a}_{j_0,k}\right)^2$ do not depend on $\widehat{\xi}_{j,n}$. Therefore, by (2.15), it holds that $T_1 = O\left(n^{-\frac{2s}{2s+1}}\right)$. For T_2 and T_3 we have that

$$T_2 = \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^2\left\{I\left(\left|\widehat{b}_{j,k}\right| > \widehat{\xi}_{j,n}\,, |b_{j,k}| > \frac{\widehat{\xi}_{j,n}}{2}\right) + I\left(\left|\widehat{b}_{j,k}\right| > \widehat{\xi}_{j,n}\,, |b_{j,k}| \leq \frac{\widehat{\xi}_{j,n}}{2}\right)\right\}\right]$$

and

$$T_3 = \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}b^2_{j,k}\left[\mathbb{P}\left(\left|\widehat{b}_{j,k}\right| \leq \widehat{\xi}_{j,n}\,, |b_{j,k}| \leq 2\widehat{\xi}_{j,n}\right) + \mathbb{P}\left(\left|\widehat{b}_{j,k}\right| \leq \widehat{\xi}_{j,n}, |b_{j,k}| > 2\widehat{\xi}_{j,n}\right)\right].$$

Note that

$$I\left(\left|\widehat{b}_{j,k}\right| > \widehat{\xi}_{j,n}\,, |b_{j,k}| \leq \frac{\widehat{\xi}_{j,n}}{2}\right) \leq I\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right), \tag{2.27}$$

$$I\left(\left|\widehat{b}_{j,k}\right| \leq \widehat{\xi}_{j,n}\,, |b_{j,k}| > 2\widehat{\xi}_{j,n}\right) \leq I\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right),$$

and if $\left|\widehat{b}_{j,k}\right| \leq \widehat{\xi}_{j,n}$, $|b_{j,k}| > 2\widehat{\xi}_{j,n}$, then $\left|\widehat{b}_{j,k}\right| \leq \frac{|b_{j,k}|}{2}$, and $\left|\widehat{b}_{j,k} - b_{j,k}\right| \geq \left|\widehat{b}_{j,k}\right| - |b_{j,k}| \geq \frac{|b_{j,k}|}{2}$. Therefore

$$b_{j,k}^2 \leq 4\left(\widehat{b}_{j,k} - b_{j,k}\right)^2. \tag{2.28}$$

Using (2.27) and (2.28), we get

$$\begin{aligned} T_2 + T_3 &\leq \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \mathbb{E}\left\{\left(b_{j,k} - \widehat{b}_{j,k}\right)^2 I\left(|b_{j,k}| > \frac{\widehat{\xi}_{j,n}}{2}\right)\right\} \\ &+ \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \mathbb{E}\left\{b_{j,k}^2 I\left(|b_{j,k}| \leq 2\widehat{\xi}_{j,n}\right)\right\} \\ &+ 5\sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \mathbb{E}\left\{\left(b_{j,k} - \widehat{b}_{j,k}\right)^2 I\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right)\right\} \\ &:= T' + T'' + T'''. \end{aligned}$$

Now we bound T'''. Using Cauchy-Schwarz inequality, one obtains

$$T''' \leq 5\sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \mathbb{E}^{\frac{1}{2}}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^4\right] \mathbb{P}^{\frac{1}{2}}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right).$$

From Lemma 2.7 we have that for any $y > 0$ the following exponential inequality holds:

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > 2\|f\|_\infty\left(\sqrt{\frac{y}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{y}{n}\right) + \frac{C_*}{\sqrt{n}}\right) \leq 2e^{-y}. \tag{2.29}$$

As in Comte [27], let

$$\Theta_{n,b} = \left\{\left|\frac{\left\|\widehat{f}_n\right\|_\infty}{\|f\|_\infty} - 1\right| < b\right\},$$

with $b \in (0,1)$. Then, using that $\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > B + D\right) \leq \mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > B\right)$ for $B, D > 0$, and taking $x_n = \frac{\delta \log n}{(1-b)^2}$, one gets

$$\begin{aligned} &\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right) \leq \mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > 2\left\|\widehat{f}_n\right\|_\infty\left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{x_n}{n}\right)\right) \\ &\leq \mathbb{P}\left(\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > 2\left\|\widehat{f}_n\right\|_\infty\left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{x_n}{n}\right)\right) \mid \Theta_{n,b}\right)\mathbb{P}\left(\Theta_{n,b}\right) \\ &+ \mathbb{P}\left(\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > 2\left\|\widehat{f}_n\right\|_\infty\left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{x_n}{n}\right)\right) \mid \Theta_{n,b}^c\right)\mathbb{P}\left(\Theta_{n,b}^c\right) \\ &:= P_1\mathbb{P}\left(\Theta_{n,b}\right) + P_2\mathbb{P}\left(\Theta_{n,b}^c\right). \end{aligned}$$

In Comte [27] is proved that if $\|f - f_n\|_\infty \leq \frac{1}{4}\|f\|_\infty$ then $\mathbb{P}\left(\Theta_{n,b}^c\right) \leq O\left(n^{-4}\right)$ for the choices of $1 \geq b \approx \frac{4}{6}\sqrt{\frac{5}{\pi}} = 0.841 \geq \frac{3}{4}$ and $N_n \leq \frac{\kappa}{(r+1)^2}\frac{n}{\log n}$, where $\kappa = \frac{1}{36}$. Following its proof it can be shown that this bound can be improved taking $\kappa = \frac{1}{36\left(\frac{7}{5}\right)}$ and b as before (see the three last equations of page 290 in [27]). These values of κ and b are the numerical constants in the hypothesis of Theorem 2.3. With this selection of κ we obtain that $\mathbb{P}\left(\Theta_{n,b}^c\right) \leq O\left(n^{-6}\right)$. Using that $\mathbb{P}\left(\Theta_{n,b}\right) = O\left(1\right)$ and $P_2 = O\left(1\right)$, it only remains to bound the conditional probability P_1. On $\Theta_{n,b}$ the following inequalities hold:

$$\left\|\widehat{f}_n\right\|_\infty > (1-b)\left\|f\right\|_\infty \tag{2.30}$$

and

$$\left\|\widehat{f}_n\right\|_\infty < (1+b)\left\|f\right\|_\infty . \tag{2.31}$$

Then, using (2.30) we get

$$P_1 \leq \mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > 2\left\|f\right\|_\infty \left(\sqrt{\frac{\delta \log n}{n}} + 2^{\frac{j}{2}}\left\|\psi\right\|_\infty \frac{\delta \log n}{n}\right)\right) \leq 2e^{-\delta \log n} = 2n^{-\delta},$$

where the last inequality is obtained using (2.29) for $y = \delta \log n > 0$. Hence, using that $\delta = 6$, we get

$$\mathbb{P}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right) \leq O\left(n^{-6}\right).$$

Therefore

$$\begin{aligned} T''' &\leq 5\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} \mathbb{E}^{\frac{1}{2}}\left[\left(b_{j,k} - \widehat{b}_{j,k}\right)^4\right]\mathbb{P}^{\frac{1}{2}}\left(\left|\widehat{b}_{j,k} - b_{j,k}\right| > \frac{\widehat{\xi}_{j,n}}{2}\right) \\ &\leq C\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} n^{-2} = Cn^{-2}\sum_{j=j_0}^{j_1} 2^j \leq O\left(n^{-2}2^{j_1}\right) \\ &= O\left(\frac{n^{-1}}{\log n}\right) \leq O\left(n^{-\frac{2s}{2s+1}}\right). \end{aligned}$$

Now we bound T'. Let j_A be the integer such that $2^{j_A} > \left(\frac{n}{\log n}\right)^{\frac{1}{2s+1}} > 2^{j_A - 1}$, then $T' = T_1' + T_2'$, where the first component is computed over the set of indices $j_0 \leq j \leq j_A$ and the second component over $j_A + 1 \leq j \leq j_1$. Hence, using Lemma 2.6 we obtain

$$\begin{aligned} T_1' &\leq \sum_{j=j_0}^{j_A}\sum_{k=0}^{2^j-1} \mathbb{E}\left(b_{j,k} - \widehat{b}_{j,k}\right)^2 \leq \frac{C}{n}\sum_{j=j_0}^{j_A} 2^j \\ &\leq O\left(\frac{2^{j_A}}{n}\right) = O\left(\left(\frac{n}{\log n}\right)^{\frac{1}{2s+1}} n^{-1}\right) \\ &\leq O\left(n^{-\frac{2s}{2s+1}}\right). \end{aligned}$$

For T'_2, one has

$$\begin{aligned}
T'_2 &= \sum_{j=j_A}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left\{\left(b_{j,k}-\widehat{b}_{j,k}\right)^2 I\left(|b_{j,k}|>\frac{\widehat{\xi}_{j,n}}{2},\Theta_{n,b}\right)\right\}\\
&+\sum_{j=j_A}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left\{\left(b_{j,k}-\widehat{b}_{j,k}\right)^2 I\left(|b_{j,k}|>\frac{\widehat{\xi}_{j,n}}{2},\Theta_{n,b}^c\right)\right\}\\
&:= T'_{2,1}+T'_{2,2}.
\end{aligned}$$

Now we bound $T'_{2,1}$. Using that on $\Theta_{n,b}$ inequality (2.30) holds and following the same procedures as in the proof of Theorem 2.2, we get

$$\begin{aligned}
T'_{2,1} &\le \sum_{j=j_A}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left(b_{j,k}-\widehat{b}_{j,k}\right)^2 I\left[|b_{j,k}|>2\left(1-b\right)\|f\|_\infty\left(\sqrt{\frac{x_n}{n}}+2^{\frac{j}{2}}\|\psi\|_\infty\frac{x_n}{n}\right)+\sqrt{\frac{\log n}{n}}\right]\\
&\le \frac{C}{n}\sum_{j=j_A}^{j_1}\sum_{k=0}^{2^j-1} I\left(|b_{j,k}|>2\left(1-b\right)\|f\|_\infty\left(\sqrt{\frac{x_n}{n}}+2^{\frac{j}{2}}\|\psi\|_\infty\frac{x_n}{n}\right)\right) \qquad (2.32)\\
&\le \frac{C}{4}\frac{1}{n}\sum_{j=j_A}^{j_1}\sum_{k=0}^{2^j-1}\frac{|b_{j,k}|^2}{\|f\|_\infty^2\left(\sqrt{\frac{\delta\log n}{n}}+2^{\frac{j}{2}}\|\psi\|_\infty\frac{\delta\log n}{(1-b)n}\right)^2}\\
&\le \frac{C\left(\|f\|_\infty\right)}{\left(\sqrt{\delta\log n}+\|\psi\|_\infty\left(1-b\right)^{-1}\delta n^{\frac{-s}{2s+1}}\left(\log n\right)^{\frac{4s+1}{4s+2}}\right)^2}2^{-2s^*j_A}\sum_{j=j_A}^{j_1}\sum_{k=0}^{2^j-1}2^{2js^*}|b_{j,k}|^2\\
&\le O\left(2^{-2s^*j_A}\right)=O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s^*}{2s+1}}\right),
\end{aligned}$$

where we have used the fact that $\sqrt{\delta\log n}+\|\psi\|_\infty\left(1-b\right)^{-1}\delta n^{\frac{-s}{2s+1}}\left(\log n\right)^{\frac{4s+1}{4s+2}}\to+\infty$ when $n\to+\infty$ and that condition (2.19) is satisfied. Now remark that if $p=2$ then $s^*=s$ and thus

$$T'_{2,1}=O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right). \tag{2.33}$$

For the case $1 \leq p < 2$, from (2.32) we have that

$$\begin{aligned}
T'_{2,1} &\leq \frac{C}{n} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} I\left(|b_{j,k}| > 2\left\|f\right\|_\infty \left(\sqrt{\frac{\delta \log n}{n}} + 2^{\frac{j}{2}} \left\|\psi\right\|_\infty \frac{\delta \log n}{(1-b)\, n}\right)\right) \\
&\leq \frac{C}{n} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} I\left(|b_{j,k}| > 2\left\|f\right\|_\infty \sqrt{\delta \frac{\log n}{n}}\right) \\
&\leq \frac{C}{n} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^{-p} \left|b_{j,k}\right|^{p} I\left(|b_{j,k}|^{-p} < \left(2\left\|f\right\|_\infty \sqrt{\delta \frac{\log n}{n}}\right)^{-p}\right) \\
&\leq C\left(\left\|f\right\|_\infty, \delta, p\right) \frac{1}{n} \left(\frac{\log n}{n}\right)^{-\frac{p}{2}} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^{p} \\
&\leq (\log n)\, C\left(\left\|f\right\|_\infty, \delta, p\right) \frac{(\log n)^{-\frac{p}{2}}}{n^{1-\frac{p}{2}}} \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^{p} = O\left(\frac{(\log n)^{1-\frac{p}{2}}}{n^{1-\frac{p}{2}}} 2^{-p j_A s^*}\right) \\
&\leq O\left(\left(\frac{n}{\log n}\right)^{\frac{p}{2}-1} \left(\frac{n}{\log n}\right)^{-\frac{p\left(s+\frac{1}{2}-\frac{1}{p}\right)}{2s+1}}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right),
\end{aligned} \tag{2.34}$$

where we have used condition (2.21). Hence $T'_{2,1} = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right)$.

Now we bound $T'_{2,2}$. Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
T'_{2,2} &\leq C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} n P^{\frac{1}{2}} \left(|b_{j,k}| > 2\left\|\widehat{f}_n\right\|_\infty \left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}} \left\|\psi\right\|_\infty \frac{x_n}{n}\right) + \sqrt{\frac{\log n}{n}} \mid \Theta^c_{n,b}\right) \mathbb{P}^{\frac{1}{2}}\left(\Theta^c_{n,b}\right) \\
&\leq C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} n \mathbb{P}^{\frac{1}{2}}\left(\Theta^c_{n,b}\right) \leq C \sum_{j=j_A}^{j_1} \sum_{k=0}^{2^j-1} n^{-2} \leq O\left(\frac{2^{j_1}}{n^2}\right) \leq O\left(\frac{n^{-1}}{\log n}\right) \leq O\left(n^{-\frac{2s}{2s+1}}\right),
\end{aligned} \tag{2.35}$$

where we have used that $\mathbb{E}\left\{\left(b_{j,k} - \widehat{b}_{j,k}\right)^4\right\} = O\left(n^2\right)$ and that $\mathbb{P}\left(\Theta^c_{n,b}\right) \leq O\left(n^{-6}\right)$. Then, putting together (2.33), (2.34) and (2.35), we obtain that $T'_2 = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right)$.

Now we bound T''. Set j_A as before, then $T'' = T''_1 + T''_2$, where the first component is calculated over the set of indices $j_0 \leq j \leq j_A$ and the second component over $j_A + 1 \leq$

$j \leq j_1$. Recall that $x_n = \frac{\delta \log n}{(1-b)^2}$, then

$$\begin{aligned}
T_1'' &= \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(|b_{j,k}| \leq 4 \left[2 \left\| \widehat{f}_n \right\|_\infty \left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}} \|\psi\|_\infty \frac{x_n}{n} \right) + \sqrt{\frac{\log n}{n}} \right] \right) \\
&\leq \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(|b_{j,k}| \leq 4 \left[2(1+b) \|f\|_\infty \left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}} \|\psi\|_\infty \frac{x_n}{n} \right) + \sqrt{\frac{\log n}{n}} \right] \right) \\
&+ \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(\Theta_{n,b}^c \right) \\
&:= T_{1,1}'' + T_{1,2}'',
\end{aligned}$$

where we have used that given $\Theta_{n,b}$ inequality (2.31) holds. Now we bound $T_{1,1}''$. For $T_{1,1}''$ we have

$$\begin{aligned}
T_{1,1}'' &= \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(|b_{j,k}|^2 \leq 16 \left[2(1+b) \|f\|_\infty \left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}} \|\psi\|_\infty \frac{x_n}{n} \right) + \sqrt{\frac{\log n}{n}} \right]^2 \right) \\
&\leq 16 \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \left[2(1+b) \|f\|_\infty \left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}} \|\psi\|_\infty \frac{x_n}{n} \right) + \sqrt{\frac{\log n}{n}} \right]^2 \\
&\leq 32 \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \left(8(1+b)^2 \|f\|_\infty^2 \left(\frac{\delta \log n}{(1-b)^2 n} + 2^j \|\psi\|_\infty^2 \frac{\delta^2 (\log n)^2}{(1-b)^4 n^2} \right) + \frac{\log n}{n} \right) \\
&\leq C \sum_{k=0}^{2^j-1} \left(8(1+b)^2 \|f\|_\infty^2 \left(\frac{\delta \log n}{(1-b)^2 n} + 2^{j_A} \|\psi\|_\infty^2 \frac{\delta^2 (\log n)^2}{(1-b)^4 n^2} \right) + \frac{\log n}{n} \right),
\end{aligned}$$

where we have used repeatedly that $(B+D)^2 \leq 2(B^2+D^2)$ for all $B, D \in \mathbb{R}$. Then,

$$\begin{aligned}
T_{1,1}'' &\leq C \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} \left(C(\|f\|_\infty, b) \left(\frac{\delta \log n}{(1-b)^2 n} + \|\psi\|_\infty^2 \frac{\delta^2 \log n}{(1-b)^4 n} \right) + \frac{\log n}{n} \right) \\
&= C(\|f\|_\infty, \|\psi\|_\infty, \delta, b) \frac{\log n}{n} \sum_{j=j_0}^{j_A} 2^j \leq C(\|f\|_\infty, \|\psi\|_\infty, \delta, b) \frac{\log n}{n} 2^{j_A} \\
&= O\left(\frac{\log n}{n} \left(\frac{n}{\log n} \right)^{\frac{1}{2s+1}} \right) = O\left(\left(\frac{n}{\log n} \right)^{-\frac{2s}{2s+1}} \right). \qquad (2.36)
\end{aligned}$$

To bound $T_{1,2}''$ we use again that $\mathbb{P}\left(\Theta_{n,b}^c\right) \leq O\left(n^{-6}\right)$ and that condition (2.19) is satisfied. Then

$$T_{1,2}'' \leq \sum_{j=j_0}^{j_A} \sum_{k=0}^{2^j-1} b_{j,k}^2 n^{-6} \leq n^{-6} \sum_{j=j_0}^{j_A} C 2^{-2js^*} = O\left(n^{-6} 2^{-2j_0 s^*}\right) \leq O\left(n^{-1}\right). \qquad (2.37)$$

Hence, by (2.36) and (2.37), $T_1'' = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right)$. Now we bound T_2''.

$$T_2'' \leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(|b_{j,k}| \leq 2\widehat{\xi}_{j,n}\right) \leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} b_{j,k}^2 = O\left(2^{-2j_A s^*}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s^*}{2s+1}}\right),$$

where we have used again the condition (2.19). Now remark that if $p = 2$ then $s^* = s$ and thus

$$T_2'' = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s^*}{2s+1}}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).$$

For $1 \leq p < 2$, we proceed as follows.

$$\begin{aligned}
T_2'' &= \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} \mathbb{E}\left[b_{j,k}^2 I\left(|b_{j,k}| \leq 2\widehat{\xi}_{j,n}\,, \Theta_{n,b}\right) + b_{j,k}^2 I\left(|b_{j,k}| \leq 2\widehat{\xi}_{j,n}\,, \Theta_{n,b}^c\right)\right] \\
&\leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} \mathbb{E}\left[b_{j,k}^2 I\left(|b_{j,k}| \leq 4\left(2\left(1+b\right)\|f\|_\infty\left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{x_n}{n}\right) + \sqrt{\frac{\log n}{n}}\right)\right)\right] \\
&+ \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(|b_{j,k}| \leq \left(8\left\|\widehat{f}_n\right\|_\infty\left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{x_n}{n}\right) + 4\sqrt{\frac{\log n}{n}}\right) \mid \Theta_{n,b}^c\right) \mathbb{P}\left(\Theta_{n,b}^c\right) \\
&:= T_{2,1}'' + T_{2,2}'',
\end{aligned}$$

where we have used that on $\Theta_{n,b}$ inequality (2.31) holds.

Now we bound $T_{2,1}''$.

$$\begin{aligned}
T_{2,1}'' &\leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^{2-p} |b_{j,k}|^p\, I\left(|b_{j,k}| \leq 8\left(1+b\right)\|f\|_\infty\left(\sqrt{\frac{x_n}{n}} + 2^{\frac{j}{2}}\|\psi\|_\infty \frac{x_n}{n}\right) + 4\sqrt{\frac{\log n}{n}}\right) \\
&\leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} \left(8\left(1+b\right)\|f\|_\infty\left(\sqrt{\frac{\delta \log n}{(1-b)^2 n}} + 2^{\frac{j_1}{2}} \frac{\|\psi\|_\infty \delta \log n}{(1-b)^2 n}\right) + 4\sqrt{\frac{\log n}{n}}\right)^{2-p} |b_{j,k}|^p \\
&\leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} \left(C\left(\|f\|_\infty, b\right)\left(\sqrt{\frac{\delta}{(1-b)^2}} + \frac{\|\psi\|_\infty \delta}{(1-b)^2}\right)\sqrt{\frac{\log n}{n}} + 4\sqrt{\frac{\log n}{n}}\right)^{2-p} |b_{j,k}|^p \\
&\leq C\left(\|f\|_\infty, b, \delta, \|\psi\|_\infty\right)^{2-p}\left(\sqrt{\frac{\log n}{n}}\right)^{2-p} \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} |b_{j,k}|^p \leq O\left(\left(\frac{\log n}{n}\right)^{\frac{2-p}{2}} 2^{-pj_A s^*}\right) \\
&= O\left(\left(\frac{\log n}{n}\right)^{\frac{2-p}{2}}\left(\frac{n}{\log n}\right)^{-\frac{p\left(s+\frac{1}{2}-\frac{1}{p}\right)}{2s+1}}\right) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right),
\end{aligned} \tag{2.38}$$

where we have used that condition (2.21) is satisfied. To bound $T_{2,2}''$ we use again that $\mathbb{P}\left(\Theta_{n,b}^c\right) \leq O\left(n^{-6}\right)$ and that condition (2.19) also holds. Then, from (2.37) we get

$$T_{2,2}'' \leq \sum_{j=j_A+1}^{j_1} \sum_{k=0}^{2^j-1} b_{j,k}^2 \mathbb{P}\left(\Theta_{n,b}^c\right) \leq n^{-6} \sum_{j=j_A+1}^{j_1} C 2^{-2js^*} = O\left(n^{-6} 2^{-2j_A s^*}\right) \leq O\left(n^{-1}\right). \tag{2.39}$$

Hence, by (2.38) and (2.39), $T'' = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right)$. Combining all terms in (2.26), we conclude that:

$$\mathbb{E}\left\|\beta - \widehat{\beta}_{\widehat{\xi}_{j,n}}\right\|_{\ell_2}^2 = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).$$

This completes the proof. □

2.5.4 Proof of Theorem 2.1

First, one needs the following proposition.

Proposition 2.1. *Let $\beta_{j,k} = \langle f, \psi_{j,k}\rangle$ and $\widehat{\beta}_{j,k} = \langle I_n, \psi_{j,k}\rangle$ with $(j,k) \in \Lambda_{j_1}$. Suppose that $f \in F^s_{p,q}(M)$ with $s > 1/p$ and $1 \leq p \leq 2$. Let $M_1 > 0$ be a constant such that $M_1^{-1} \leq f \leq M_1$ (see Lemma 2.5). Let $\epsilon_{j_1} = 2M_1^2 e^{2\gamma_{j_1}+1} D_{j_1} A_{j_1}$. If $\epsilon_{j_1} \leq 1$, then there exists $\theta^*_{j_1} \in \mathbb{R}^{|\Lambda_{j_1}|}$ such that:*

$$\left\langle f_{j_1,\theta^*_{j_1}}, \psi_{j,k}\right\rangle = \langle f, \psi_{j,k}\rangle = \beta_{j,k} \text{ for all } (j,k) \in \Lambda_{j_1}.$$

Moreover, the following inequality holds (approximation error)

$$\Delta\left(f; f_{j_1,\theta^*_{j_1}}\right) \leq \frac{M_1}{2} e^{\gamma_{j_1}} D^2_{j_1}.$$

Suppose that Assumptions 2.1 and 2.2 hold. Let $\eta_{j_1,n} = 4M_1^2 e^{2\gamma_{j_1}+2\epsilon_{j_1}+2} A^2_{j_1} \frac{|\Lambda_{j_1}|}{n}$. Then, for every $\lambda > 0$ such that $\lambda \leq \eta^{-1}_{j_1,n}$ there exists a set $\Omega_{n,1}$ of probability less than $M_2\lambda^{-1}$, where M_2 is the constant defined in Lemma 2.6, such that outside the set $\Omega_{n,1}$ there exists some $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$ which satisfies:

$$\left\langle f_{j_1,\widehat{\theta}_n}, \psi_{j,k}\right\rangle = \langle I_n, \psi_{j,k}\rangle = \widehat{\beta}_{j,k} \text{ for all } (j,k) \in \Lambda_{j_1}.$$

Moreover, outside the set $\Omega_{n,1}$, the following inequality holds (estimation error)

$$\Delta\left(f_{j_1,\theta^*_{j_1}}; f_{j_1,\widehat{\theta}_n}\right) \leq 2M_1 e^{\gamma_{j_1}+\epsilon_{j_1}+1} M_2 \lambda \frac{|\Lambda_{j_1}|}{n}.$$

Proof of Proposition 2.1.

Approximation error: Recall that $\beta_{j,k} = \langle f, \psi_{j,k}\rangle$ and let $\beta = (\beta_{j,k})_{(j,k)\in\Lambda_{j_1}}$. Define by $g_{j_1} = \sum_{(j,k)\in\Lambda_{j_1}} \theta_{j,k}\psi_{j,k}$ an approximation of $g = \log(f)$ and let $\beta_{0,(j,k)} = \left\langle f_{j_1,\theta_{j_1}}, \psi_{j,k}\right\rangle = \langle \exp(g_{j_1}), \psi_{j,k}\rangle$ with $\theta_{j_1} = (\theta_{j,k})_{(j,k)\in\Lambda_{j_1}}$ and $\beta_0 = \left(\beta_{0,(j,k)}\right)_{(j,k)\in\Lambda_{j_1}}$. Observe that the coefficients $\beta_{j,k} - \beta_{0,(j,k)}$, $(j,k) \in \Lambda_{j_1}$, are the coefficients of the orthonormal projection of $f - f_{j_1,\theta_{j_1}}$ onto V_j. Hence, by Bessel's inequality, $\|\beta - \beta_0\|^2_{\ell_2} \leq \left\|f - f_{j_1,\theta_{j_1}}\right\|^2_{L_2}$. Using

Lemma 2.5 and Lemma 2 in Barron and Sheu [9], we get that:

$$\begin{aligned}
\|\beta-\beta_0\|_{\ell_2}^2 \le \int \left(f - f_{j_1,\theta_{j_1}}\right)^2 d\mu &\le M_1 \int \frac{\left(f-f_{j_1,\theta_{j_1}}\right)^2}{f} d\mu \\
&\le M_1 e^{2\left\|\log\left(\frac{f}{f_{j_1,\theta_{j_1}}}\right)\right\|_\infty} \int f \left(\log\left(\frac{f}{f_{j_1,\theta_{j_1}}}\right)\right)^2 d\mu \\
&\le M_1^2 e^{2\|g-g_{j_1}\|_\infty} \|g-g_{j_1}\|_{L_2}^2 = M_1^2 e^{2\gamma_{j_1}} D_{j_1}^2.
\end{aligned}$$

Then, one can easily check that $b = e^{\left(\left\|\log\left(f_{j_1,\theta_{j_1}}\right)\right\|_\infty\right)} \le M_1 e^{\gamma_{j_1}}$. Thus, the assumption that $\epsilon_{j_1} \le 1$ implies that the following inequality $\|\beta-\beta_0\|_{\ell_2} \le M_1 e^{\gamma_{j_1}} D_{j_1} \le \frac{1}{2beA_{j_1}}$ is satisfied. Hence, Lemma 2.4 can be applied with $\theta_0 = \theta_{j_1}$, $\widetilde{\beta} = \beta$ and $b = \exp\left(\left\|\log\left(f_{j_1,\theta_{j_1}}\right)\right\|_\infty\right)$, which implies that there exists $\theta_{j_1}^* = \theta(\beta) \in \mathbb{R}^{|\Lambda_{j_1}|}$ such that $\left\langle f_{j_1,\theta_{j_1}^*}, \psi_{j,k}\right\rangle = \beta_{j,k}$ for all $(j,k) \in \Lambda_{j_1}$.

By the Pythagorian-like relationship (2.9), we obtain that $\Delta\left(f; f_{j_1,\theta_{j_1}^*}\right) \le \Delta\left(f; f_{j_1,\theta_{j_1}}\right)$. Hence, using Lemma 2.2, it follows that

$$\begin{aligned}
\Delta\left(f; f_{j_1,\theta_{j_1}^*}\right) &\le \frac{1}{2} e^{\left\|\log\left(\frac{f}{f_{j_1,\theta_{j_1}}}\right)\right\|_\infty} \int f \left(\log\left(\frac{f}{f_{j_1,\theta_{j_1}}}\right)\right)^2 d\mu \\
&\le \frac{M_1}{2} e^{\left\|\log(f)-\log\left(f_{j_1,\theta_{j_1}}\right)\right\|_\infty} \int \left(\log(f) - \log\left(f_{j_1,\theta_{j_1}}\right)\right)^2 d\mu \\
&= \frac{M_1}{2} e^{\|g-g_{j_1}\|_\infty} \|g-g_{j_1}\|_{L_2}^2 = \frac{M_1}{2} e^{\gamma_{j_1}} D_{j_1}^2,
\end{aligned}$$

which completes the proof for the approximation error.

Estimation error: Applying Lemma 2.4 with $\theta_0 = \theta_{j_1}^*$, $\beta_{0,(j,k)} = \langle f_{j_1,\theta_0}, \psi_{j,k}\rangle = \beta_{j,k}$, $\widetilde{\beta} = \widehat{\beta}$, where $\widehat{\beta} = \left(\widehat{\beta}_{j,k}\right)_{(j,k)\in\Lambda_{j_1}}$, and $b = \exp\left(\left\|\log\left(f_{j_1,\theta_{j_1}^*}\right)\right\|_\infty\right)$ we obtain that if $\left\|\widehat{\beta}-\beta\right\|_{\ell_2} \le \frac{1}{2ebA_{j_1}}$ with $\beta = (\beta_{j,k})_{(j,k)\in\Lambda_{j_1}}$ then there exists $\widehat{\theta}_n = \theta\left(\widehat{\beta}\right)$ such that $\left\langle f_{j_1,\widehat{\theta}_n}, \psi_{j,k}\right\rangle = \widehat{\beta}_{j,k}$ for all $(j,k) \in \Lambda_{j_1}$.

Hence, it remains to prove that our assumptions imply that $\left\|\widehat{\beta}-\beta\right\|_{\ell_2} \le \frac{1}{2ebA_{j_1}}$ holds with probability $1 - M_2\lambda^{-1}$. First remark that $b \le M_1 e^{\gamma_{j_1}+\epsilon_{j_1}}$ and that by Markov's inequality and Lemma 2.6 we obtain that for any $\lambda > 0$,

$$\mathbb{P}\left(\left\|\widehat{\beta}-\beta\right\|_{\ell_2}^2 \ge \lambda \frac{|\Lambda_{j_1}|}{n}\right) \le \frac{1}{\lambda} \frac{n}{|\Lambda_{j_1}|} \mathbb{E}\left\|\widehat{\beta}-\beta\right\|_{\ell_2}^2 \le M_2\lambda^{-1}.$$

Hence, outside a set $\Omega_{n,1}$ of probability less than $M_2\lambda^{-1}$ then $\left\|\widehat{\beta}-\beta\right\|_{\ell_2}^2 \le \lambda\frac{|\Lambda_{j_1}|}{n}$. Therefore, the condition $\left\|\widehat{\beta}-\beta\right\|_{\ell_2} \le \frac{1}{2ebA_{j_1}}$ holds if $\left(\lambda\frac{|\Lambda_{j_1}|}{n}\right)^{\frac{1}{2}} \le \frac{1}{2ebA_{j_1}}$, which is equivalent to $4e^2 b^2 A_{j_1}^2 \lambda \frac{|\Lambda_{j_1}|}{n} \leq 1$. Using that $b^2 \leq M_1^2 e^{2\gamma_{j_1}+2\epsilon_{j_1}}$ the last inequality is true if $\eta_{j_1,n} = 4M_1^2 e^{2\gamma_{j_1}+2\epsilon_{j_1}+2} A_{j_1}^2 \frac{|\Lambda_{j_1}|}{n} \leq \frac{1}{\lambda}$.

Hence, outside the set $\Omega_{n,1}$, our assumptions imply that there exists $\widehat{\theta}_n = \theta\left(\widehat{\beta}\right)$ such that $\left\langle f_{j_1,\widehat{\theta}_n}, \psi_{j,k} \right\rangle = \widehat{\beta}_{j,k}$ for all $(j,k) \in \Lambda_{j_1}$. Finally, outside the set $\Omega_{n,1}$, by using the bound given in Lemma 2.4, one obtains the following inequality for the estimation error

$$\Delta\left(f_{j_1,\theta^*_{j_1}}; f_{j_1,\widehat{\theta}_n}\right) \leq 2M_1 e^{\gamma_{j_1}+\epsilon_{j_1}+1} \lambda \frac{|\Lambda_{j_1}|}{n},$$

which completes the proof of Proposition 2.1.

Our assumptions on $j_1(n)$ imply that $\frac{1}{2} n^{\frac{1}{2s+1}} \leq 2^{j_1(n)} \leq n^{\frac{1}{2s+1}}$. Therefore, using Lemma 2.1, one has that for all $f \in F_{2,2}^s(M)$ with $s > 1/2$

$$\gamma_{j_1(n)} \leq C n^{\frac{1-2s}{2(2s+1)}}, \quad A_{j_1(n)} \leq C n^{\frac{1}{2(2s+1)}}, \quad D_{j_1(n)} \leq C n^{-\frac{s}{2s+1}},$$

where C denotes constants not depending on $g = \log(f)$. Hence,

$$\lim_{n\to+\infty} \epsilon_{j_1(n)} = \lim_{n\to+\infty} 2M_1^2 e^{2\gamma_{j_1(n)}+1} A_{j_1(n)} D_{j_1(n)} = 0$$

uniformly over $F_{2,2}^s(M)$ for $s > 1/2$. For all sufficiently large n, $\epsilon_{j_1(n)} \leq 1$ and thus, using Proposition 2.1, there exists $\theta^*_{j_1(n)} \in \mathbb{R}^{|\Lambda_{j_1(n)}|}$ such that

$$\Delta\left(f; f_{j,\theta^*_{j_1(n)}}\right) \leq \frac{M_1}{2} e^{\gamma_{j_1(n)}} D_{j_1(n)}^2 \leq C n^{-\frac{2s}{2s+1}} \text{ for all } f \in F_{2,2}^s(M). \tag{2.40}$$

By the same arguments it follows that

$$\lim_{n\to+\infty} \eta_{j_1(n),n} = \lim_{n\to+\infty} 4M_1^2 e^{2\gamma_{j_1(n)}+2\epsilon_{j_1(n)}+2} A_{j_1(n)}^2 \frac{|\Lambda_{j_1(n)}|}{n} = 0$$

uniformly over $F_{2,2}^s(M)$ for $s > 1/2$. Now let $\lambda > 0$. The above result shows that for sufficiently large n, $\lambda \leq \eta_{j_1(n),n}^{-1}$, and thus using Proposition 2.1 it follows that there exists a set $\Omega_{n,1}$ of probability less than $M_2\lambda^{-1}$ such that outside this set there exists $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1(n)}|}$ which satisfies:

$$\Delta\left(f_{j_1(n),\theta^*_{j_1(n)}}; f_{j_1(n),\widehat{\theta}_n}\right) \leq 2M_1 e^{\gamma_{j_1(n)}+\epsilon_{j_1(n)}+1} M_2 \lambda \frac{|\Lambda_{j_1(n)}|}{n} \leq C\lambda n^{-\frac{2s}{2s+1}}, \tag{2.41}$$

for all $f \in F_{2,2}^s(M)$. Then, by the Pythagorian-like identity (2.9) it follows that outside the set $\Omega_{n,1}$

$$\Delta\left(f; f_{j_1(n),\widehat{\theta}_n}\right) = \Delta\left(f; f_{j_1(n),\theta^*_{j_1(n)}}\right) + \Delta\left(f_{j_1(n),\theta^*_{j_1(n)}}; f_{j_1(n),\widehat{\theta}_n}\right),$$

and thus Theorem 2.1 follows from inequalities (2.40) and (2.41). □

2.5.5 Proof of Theorem 2.2

First, one needs the following proposition.

Proposition 2.2. *Let* $\beta_{j,k} := \langle f, \psi_{j,k}\rangle$ *and* $\widehat{\beta}_{\xi_{j,n},(j,k)} := \delta_{\xi_{j,n}}\left(\widehat{\beta}_{j,k}\right)$ *with* $(j,k) \in \Lambda_{j_1}$*. Assume that* $f \in F^s_{p,q}(M)$ *with* $s > 1/p$ *and* $1 \leq p \leq 2$*. Let* $M_1 > 0$ *be a constant such that* $M_1^{-1} \leq f \leq M_1$ *(see Lemma 2.5). Let* $\epsilon_{j_1} = 2M_1^2 e^{2\gamma_{j_1}+1} D_{j_1} A_{j_1}$*. If* $\epsilon_{j_1} \leq 1$*, then there exists* $\theta^*_{j_1} \in \mathbb{R}^{|\Lambda_{j_1}|}$ *such that:*

$$\left\langle f_{j_1,\theta^*_{j_1}}, \psi_{j,k}\right\rangle = \langle f, \psi_{j,k}\rangle = \beta_{j,k} \text{ for all } (j,k) \in \Lambda_{j_1}.$$

Moreover, the following inequality holds (approximation error)

$$\Delta\left(f; f_{j_1,\theta^*_{j_1}}\right) \leq \frac{M_1}{2} e^{\gamma_{j_1}} D^2_{j_1}.$$

Suppose that Assumptions 2.1 and 2.2 hold. Let $\eta_{j_1,n} = 4M_1^2 e^{2\gamma_{j_1}+2\epsilon_{j_1}+2} A^2_{j_1}\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}$*. Then, for every* $\lambda > 0$ *such that* $\lambda \leq \eta^{-1}_{j_1,n}$ *there exists a set* $\Omega_{n,2}$ *of probability less than* $M_3\lambda^{-1}$*, where* M_3 *is the constant defined in Lemma 2.8, such that outside the set* $\Omega_{n,2}$ *there exists some* $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1}|}$ *which satisfies:*

$$\left\langle f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}, \psi_{j,k}\right\rangle = \delta_{\xi_{j,n}}\left(\widehat{\beta}_{j,k}\right) = \widehat{\beta}_{\xi_{j,n},(j,k)} \text{ for all } (j,k) \in \Lambda_{j_1}.$$

Moreover, outside the set $\Omega_{n,2}$*, the following inequality holds (estimation error)*

$$\Delta\left(f_{j_1,\theta^*_{j_1}}; f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}\right) \leq 2M_1 e^{\gamma_{j_1}+\epsilon_{j_1}+1}\lambda\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}.$$

Proof of Proposition 2.2.

Approximation error: The proof is the same that the one of Proposition 2.1.

Estimation error: Applying Lemma 2.4 with $\theta_0 = \theta^*_{j_1}$, $\beta_{0,(j,k)} = \langle f_{j_1,\theta_0}, \psi_{j,k}\rangle = \beta_{j,k}$, $\widetilde{\beta} = \widehat{\beta}_{\xi_{j,n}}$, where $\widehat{\beta}_{\xi_{j,n}} = \left(\widehat{\beta}_{\xi_{j,n},(j,k)}\right)_{(j,k)\in\Lambda_{j_1}}$, and $b = \exp\left(\left\|\log\left(f_{j_1,\theta^*_{j_1}}\right)\right\|_\infty\right)$ we obtain that if $\left\|\widehat{\beta}_{\xi_{j,n}} - \beta\right\|_{\ell_2} \leq \frac{1}{2ebA_{j_1}}$ with $\beta = (\beta_{j,k})_{(j,k)\in\Lambda_{j_1}}$ then there exists $\widehat{\theta}_n = \theta\left(\widehat{\beta}_{\xi_{j,n}}\right)$ such that $\left\langle f^{HT}_{j_1,\widehat{\theta}_{j_1},\xi_{j,n}}, \psi_{j,k}\right\rangle = \widehat{\beta}_{\xi_{j,n},(j,k)}$ for all $(j,k) \in \Lambda_{j_1}$.

Hence, it remains to prove that our assumptions imply that $\left\|\widehat{\beta}_{\xi_{j,n}} - \beta\right\|_{\ell_2} \leq \frac{1}{2ebA_{j_1}}$ holds with probability $1 - M_3\lambda^{-1}$. First remark that $b \leq M_1 e^{\gamma_{j_1}+\epsilon_{j_1}}$ and that by Markov's inequality and Lemma 2.8 we obtain that for any $\lambda > 0$,

$$\begin{aligned}\mathbb{P}\left(\left\|\widehat{\beta}_{\xi_{j,n}} - \beta\right\|^2_{\ell_2} \geq \lambda\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right) &\leq \frac{1}{\lambda}\left(\frac{n}{\log n}\right)^{\frac{2s}{2s+1}} \mathbb{E}\left\|\widehat{\beta}_{\xi_{j,n}} - \beta\right\|^2_{\ell_2} \\ &\leq \frac{M_3}{\lambda}\left(\frac{n}{\log n}\right)^{\frac{2s}{2s+1}}\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}} \leq M_3\lambda^{-1}.\end{aligned}$$

Therefore, inequality $\left\| \widehat{\beta}_{\xi_{j,n}} - \beta \right\|_{\ell_2}^2 \leq \lambda \left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}$ holds outside a set $\Omega_{n,2}$ of probability less than $M_3\lambda^{-1}$. Hence, the condition $\left\| \widehat{\beta}_{\xi_{j,n}} - \beta \right\|_{\ell_2} \leq \frac{1}{2ebA_{j_1}}$ is satisfied if $\left(\lambda \left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right)^{\frac{1}{2}} \leq \frac{1}{2ebA_{j_1}}$, which is equivalent to $4e^2b^2A_{j_1}^2\lambda\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}} \leq 1$. Using that $b^2 \leq M_1^2 e^{2\gamma_{j_1}+2\epsilon_{j_1}}$ the last inequality is true if $\eta_{j_1,n} = 4M_1^2 e^{2\gamma_{j_1}+2\epsilon_{j_1}+2} A_{j_1}^2 \left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}} \leq \frac{1}{\lambda}$.

Thus, outside the set $\Omega_{n,2}$, our assumptions imply that there exists $\widehat{\theta}_n = \theta\left(\widehat{\beta}_{\xi_{j,n}}\right)$ such that $\left\langle f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}, \psi_{j,k}\right\rangle = \widehat{\beta}_{\xi_{j,n},(j,k)}$ for all $(j,k) \in \Lambda_{j_1}$. Finally, outside the set $\Omega_{n,2}$, by using the bound given in Lemma 2.4, one obtains the following inequality for the estimation error

$$\Delta\left(f_{j_1,\theta^*_{j_1}}; f^{HT}_{j_1,\widehat{\theta}_n,\xi_{j,n}}\right) \leq 2M_1 e^{\gamma_{j_1}+\epsilon_{j_1}+1}\lambda\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}},$$

which completes the proof of Proposition 2.2.

Our assumptions on $j_1(n)$ imply that $\frac{1}{2}\frac{n}{\log n} \leq 2^{j_1(n)} \leq \frac{n}{\log n}$. Therefore, using Lemma 2.1, one has that for all $f \in F^s_{p,q}(M)$ with $s > 1/p$

$$\gamma_{j_1(n)} \leq C\left(\frac{n}{\log n}\right)^{-\left(s-\frac{1}{p}\right)}, \quad A_{j_1(n)} \leq \left(\frac{n}{\log n}\right)^{\frac{1}{2}}, \quad D_{j_1(n)} \leq C\left(\frac{n}{\log n}\right)^{-s^*},$$

where C denotes constants not depending on $g = \log(f)$. Hence,

$$\lim_{n\to+\infty} \epsilon_{j_1(n)} = \lim_{n\to+\infty} 2M_1^2 e^{2\gamma_{j_1(n)}+1} A_{j_1(n)} D_{j_1(n)} = 0$$

uniformly over $F^s_{p,q}(M)$ for $s > 1/p$. For all sufficiently large n, $\epsilon_{j_1(n)} \leq 1$ and thus, using Proposition 2.2, there exists $\theta^*_{j_1(n)} \in \mathbb{R}^{\left|\Lambda_{j_1(n)}\right|}$ such that

$$\Delta\left(f; f_{j_1(n),\theta^*_{j_1(n)}}\right) \leq \frac{M_1}{2} e^{\gamma_{j_1(n)}} D^2_{j_1(n)} \leq C\left(\frac{n}{\log n}\right)^{-2s^*} \quad \text{for all } f \in F^s_{p,q}(M).$$

Now remark that if $p = 2$ then $s^* = s > 1$ (by assumption), thus

$$\Delta\left(f; f_{j_1(n),\theta^*_{j_1(n)}}\right) = O\left(\left(\frac{n}{\log n}\right)^{-2s}\right) \leq O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right).$$

If $1 \leq p < 2$ then one can check that condition $s > \frac{1}{2} + \frac{1}{p}$ implies that $2s^* > \frac{2s}{2s+1}$, hence

$$\Delta\left(f; f_{j_1(n),\theta^*_{j_1(n)}}\right) \leq O\left(\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}\right). \tag{2.42}$$

By the same arguments it follows that

$$\lim_{n\to+\infty} \eta_{j_1(n),n} = \lim_{n\to+\infty} 4M_1^2 e^{2\left(\gamma_{j_1(n)}+\epsilon_{j_1(n)}+1\right)} A^2_{j_1(n)} \left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}} = 0$$

uniformly over $F^s_{p,q}(M)$ for $s > 1/p$. Now let $\lambda > 0$. The above result shows that for sufficiently large n, $\lambda \leq \eta^{-1}_{j_1(n),n}$. Thus, using Proposition 2.2 it follows that there exists a set $\Omega_{n,2}$ of probability less than $M_3 \lambda^{-1}$ such that outside this set there exists $\widehat{\theta}_n \in \mathbb{R}^{|\Lambda_{j_1(n)}|}$ which satisfies:

$$\Delta\left(f_{j_1(n),\theta^*_{j_1(n)}}; f^{HT}_{j_1(n),\widehat{\theta}_n,\xi_{j,n}}\right) \leq 2M_1 e^{\gamma_{j_1(n)}+\epsilon_{j_1(n)}+1} \lambda \left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}} \tag{2.43}$$

for all $f \in F^s_{p,q}(M)$. Then, by the Pythagorian-like identity (2.9) it follows that outside the set $\Omega_{n,2}$

$$\Delta\left(f; f^{HT}_{j_1(n),\widehat{\theta}_n,\xi_{j,n}}\right) = \Delta\left(f; f_{j_1(n),\theta^*_{j_1(n)}}\right) + \Delta\left(f_{j_1(n),\theta^*_{j_1(n)}}; f^{HT}_{j_1(n),\widehat{\theta}_n,\xi_{j,n}}\right),$$

and thus Theorem 2.2 follows from inequalities (2.42) and (2.43). □

2.5.6 Proof of Theorem 2.3

The proof is analogous to the one of Theorem 2.2. It follows from Lemma 2.9.

Chapter 3

Nonparametric estimation of covariance functions by model selection

Abstract: We propose a model selection approach for covariance estimation of a stochastic process. Under very general assumptions, observing i.i.d. replications of the process at fixed observation points, we construct an estimator of the covariance function by expanding the process onto a collection of basis functions. This is based on the formulation of the covariance estimation problem through linear matrix regression models. We study the non asymptotic, finite sample properties of this estimate and give a tractable way of selecting the best estimator among a possible set of candidates. The optimality of the procedure is proved via an oracle inequality which warrants that the best model is selected. Some numerical experiments show the good performance of the estimator proposed.

3.1 Introduction

Estimating the covariance function of stochastic process is a fundamental issue with many applications (we refer to Stein [86], Journel [51] or Cressie [28] for general references for applications). While parametric methods have been extensively studied in the statistical literature (see Cressie [28] for a review), nonparametric procedures have only recently received a growing attention. One of the main difficulty in this framework is to impose that the estimator is also a covariance function, preventing the direct use of usual nonparametric statistical methods. In this chapter, we propose to use a model selection procedure to construct a nonparametric estimator of the covariance function of a stochastic process under general assumptions for the process. In particular we will not assume Gaussianity nor stationarity.

Consider a stochastic process $X(t)$ with values in $\mathbb{R}$, indexed by $t \in T$, a subset of $\mathbb{R}^d$, $d \in \mathbb{N}$. Throughout the chapter, we assume that its covariance function is finite, i.e. $|\sigma(s,t)| = |Cov(X(s), X(t))| < +\infty$ for all $s, t \in T$ and, for sake of simplicity, zero mean

$\mathbb{E}\left(X\left(t\right)\right)=0$ for all $t\in T$. The observations are $X_i\left(t_j\right)$ for $i=1,...,N$ and $j=1,...,n$, where the points $t_1,...,t_n\in T$ are fixed, and $X_1,...,X_N$ are independent copies of the process X.

Functional approximations of the processes X_1,...,X_N from data $(X_i(t_j))$ are involved in covariance function estimation. When dealing with functional data analysis (see, e.g., Ramsay and Silverman [76]), smoothing the processes X_1,...,X_N is sometimes carried out as a first step before computing the empirical covariance such as spline interpolation for example (see for instance Elogne, Perrin and Thomas-Agnan [40]) or projection onto a general finite basis. Let $\mathbf{x}_i=\left(X_i\left(t_1\right),...,X_i\left(t_n\right)\right)^\top$ be the vector of observations at the points $t_1,...,t_n$ with $i\in\{1,...,N\}$. Let $\{g_\lambda\}_\lambda$ be a collection of possibly independent functions $g_\lambda:T\to\mathbb{R}$, and define $\mathcal{M}$ as a generic countable set given by $\mathcal{M}=\{m: m \text{ is a set of indices }\}$. Then, let $m\in\mathcal{M}$ be a subset of indices of size $|m|\in\mathbb{N}$ and define the $n\times|m|$ matrix $\mathbf{G}$ with entries $g_{j\lambda}=g_\lambda\left(t_j\right)$, $j=1,...,n$, $\lambda\in m$. $\mathbf{G}$ will be called the design matrix corresponding to the set of basis functions indexed by m.

In such setting, usual covariance estimation is a two-step procedure: first, for each $i=1,...,N$, fit the regression model

$$\mathbf{x}_i=\mathbf{G}\mathbf{a}_i+\epsilon_i \tag{3.1}$$

(by least squares or regularized least squares), where ϵ_i are random vectors in $\mathbb{R}^n$, to obtain estimates $\widehat{\mathbf{a}}_i=(\widehat{a}_{i,\lambda})_{\lambda\in m}\in\mathbb{R}^{|m|}$ of $\mathbf{a}_i$ where in the case of standard least squares estimation (assuming for simplicity that $\mathbf{G}^\top\mathbf{G}$ is invertible)

$$\widehat{\mathbf{a}}_i=(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{x}_i, i=1,\ldots,N.$$

Then, the estimation of the covariance is obtained by computing the following matrix

$$\widehat{\mathbf{\Sigma}}=\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top, \tag{3.2}$$

where

$$\widehat{\mathbf{\Psi}}=\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{a}}_i\widehat{\mathbf{a}}_i^\top=(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}. \tag{3.3}$$

This corresponds to approximate the process X by a truncated process $\widetilde{X}_i$ defined as

$$\widetilde{X}_i\left(t\right)=\sum_{\lambda\in m}\widehat{a}_{i,\lambda}g_\lambda\left(t\right), i=1,\ldots,N,$$

and to choose the empirical covariance of $\widetilde{X}$ as an estimator of the covariance of X, defined by

$$\widehat{\sigma}\left(s,t\right)=\frac{1}{N}\sum_{i=1}^{N}\widetilde{X_i}\left(s\right)\widetilde{X_i}\left(t\right).$$

We consider the estimator (3.2) as the least squares estimator of the following matrix regression model

$$\mathbf{x}_i\mathbf{x}_i^\top=\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top+\mathbf{U}_i, \quad i=1,...,N, \tag{3.4}$$

where $\mathbf{\Psi}$ is a symmetric matrix and $\mathbf{U}_i$ are i.i.d. matrix errors. Fitting the models (3.1) and (3.4) by least squares naturally leads to the definition of different contrast and risk functions as the estimation is not performed in the same space ($\mathbb{R}^{|m|}$ for model (3.1) and $\mathbb{R}^{|m|\times|m|}$ for model (3.4)). By choosing an appropriate loss function, least squares estimation in model (3.4) also leads to the natural estimate (3.2) derived from least squares estimation in model (3.1). A similar estimate can be found in Fan, Fan and Lv [41]. However, in this chapter, we tackle the problem of model selection, i.e. choosing an appropriate data-based subset of indices $m \in \mathcal{M}$, which is very distinct in model (3.1) and model (3.4). Indeed, model selection for (3.1) depends on the variability of the vectors $\mathbf{x}_i$'s while for (3.4) it depends on the variability of the matrices $\mathbf{x}_i\mathbf{x}_i^\top$'s. One of the main contributions of this chapter is to show that considering model (3.4) enables to handle a large variety of cases and to build an optimal model selection estimator of the covariance without too strong assumptions on the model. Moreover it will be shown that considering model (3.4) leads to the estimator $\widehat{\mathbf{\Psi}}$ defined in (3.3) which lies in the class of non-negative definite matrices and thus provides a proper covariance matrix $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$.

A similar method has been developed for smooth interpolation of covariance functions in Biscay et al. [19], but restricted to basis functions that are determined by reproducing kernels in suitable Hilbert spaces and a different fitting criterion. Similar ideas are also tackled in Matsuo et al. [67]. These authors deal with the estimation of $\mathbf{\Sigma}$ within the covariance class $\mathbf{\Gamma} = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$ induced by an orthogonal wavelet expansion. However, their fitting criterion is not general since they choose the Gaussian likelihood as a contrast function, and thus their method requires specific distributional assumptions. We also point out that computation of the Gaussian likelihood requires inversion of $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$, which is not directly feasible if $rk\,(\mathbf{G}) < n$ or some diagonal entities of the non-negative definite matrix $\mathbf{\Psi}$ are zero.

Hence, to our knowledge, no previous work has proposed to use the matrix regression model (3.4) under general moments assumptions of the process X using a general basis expansion for nonparametric covariance function estimation. We point out that the asymptotic behaviour will be taken with respect to the number of replications N while the observation points t_i, $i = 1, \ldots, n$ remain fixed.

The chapter falls into the following parts. The description of the statistical framework of the matrix regression is given in Section 3.2. Section 3.3 is devoted to the main statistical results. Namely we study the behavior of the estimator for a fixed model in Section 3.3.1 while Section 3.3.2 deals with the model selection procedure and provide the oracle inequality. Section 3.4 states a concentration inequality that is used in the proofs of the main results of the chapter, while some numerical experiments are described in Section 3.5. The proofs are postponed to the end of the chapter in Section 3.6.

3.2 Nonparametric model selection for covariance estimation

Recall that $X = (X\,(t) : t \in T)$ is a real valued stochastic process, where T denotes some subset of $\mathbb{R}^d$, with $d \in \mathbb{N}$. Assume that X has finite moments up to order 4, and zero mean, i.e. $\mathbb{E}\left(X\left(t\right)\right)=0$ for all $t\in T$. The covariance function of X is denoted by $\sigma\left(s,t\right)=Cov\left(X\left(s\right),X\left(t\right)\right)$ for $s,t\in T$ and recall that $X_1,...,X_N$ are independent copies of the process X.

In this work, we observe at different points $t_1,...,t_n\in T$ independent copies of the process, denoted by $X_i\left(t_j\right)$, with $i=1,...,N$, $j=1,...,n$. Note that $\mathbf{x}_i=\left(X_i\left(t_1\right),...,X_i\left(t_n\right)\right)^{\top}$ is the vector of observations at the points $t_1,...,t_n$ for each $i=1,...,N$. The matrix $\mathbf{\Sigma}=\mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i^{\top}\right)=\left(\sigma\left(t_j,t_k\right)\right)_{1\leq j\leq n,1\leq k\leq n}$ is the covariance matrix of X at the observations points. Let $\mathbf{S}$ denote the sample covariance matrix (non corrected by the mean) of the data $\mathbf{x}_1,...,\mathbf{x}_N$, i.e.

$$\mathbf{S}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\top}.$$

Our aim is to build a model selection estimator of the covariance of the process observed with N replications but without additional assumptions such as stationarity nor Gaussianity. The asymptotics will be taken with respect to N, the number of copies of the process.

3.2.1 Notations and preliminary definitions

First, define specific matrix notations. We refer to Seber [84] or Kollo and von Rosen [54] for definitions and properties of matrix operations and special matrices. As usual, vectors in $\mathbb{R}^k$ are regarded as column vectors for all $k\in\mathbb{N}$. We will consider matrix data as a natural extension of the vectorial data, with different correlation structure. For any random matrix $\mathbf{Z}=\left(Z_{ij}\right)_{1\leq i\leq k,1\leq j\leq n}$, its expectation is denoted by $\mathbb{E}\left(\mathbf{Z}\right)=\left(\mathbb{E}\left(Z_{ij}\right)\right)_{1\leq i\leq k,1\leq j\leq n}$. For any random vector $\mathbf{z}=\left(Z_i\right)_{1\leq i\leq k}$, let $\mathbb{V}\left(\mathbf{z}\right)=\left(Cov\left(Z_i,Z_j\right)\right)_{1\leq i,j\leq k}$ be its covariance matrix. With this notation, $\mathbb{V}\left(\mathbf{x}_1\right)=\mathbb{V}\left(\mathbf{x}_i\right)=\left(\sigma\left(t_j,t_k\right)\right)_{1\leq j\leq n,1\leq k\leq n}$ is the covariance matrix of X.

Let $m\in\mathcal{M}$, and recall that to the finite set $\mathcal{G}_m=\left\{g_\lambda\right\}_{\lambda\in m}$ of functions $g_\lambda:T\to\mathbb{R}$ we associate the $n\times|m|$ matrix $\mathbf{G}$ with entries $g_{j\lambda}=g_\lambda\left(t_j\right)$, $j=1,...,n$, $\lambda\in m$. Furthermore, for each $t\in T$, we write $\mathbf{G}_t=\left(g_\lambda\left(t\right),\lambda\in m\right)^{\top}$. For $k\in\mathbb{N}$, $\mathcal{S}_k$ denotes the linear subspace of $\mathbb{R}^{k\times k}$ composed of symmetric matrices. For $\mathbf{G}\in\mathbb{R}^{n\times|m|}$, $\mathcal{S}\left(\mathbf{G}\right)$ is the linear subspace of $\mathbb{R}^{n\times n}$ defined by

$$\mathcal{S}\left(\mathbf{G}\right)=\left\{\mathbf{G\Psi G}^{\top}:\mathbf{\Psi}\in\mathcal{S}_{|m|}\right\}.$$

Let $\mathcal{S}_N\left(\mathbf{G}\right)$ be the linear subspace of $\mathbb{R}^{nN\times n}$ defined by

$$\mathcal{S}_N\left(\mathbf{G}\right)=\left\{\mathbf{1}_N\otimes\mathbf{G\Psi G}^{\top}:\mathbf{\Psi}\in\mathcal{S}_{|m|}\right\}=\left\{\mathbf{1}_N\otimes\mathbf{\Gamma}:\mathbf{\Gamma}\in\mathcal{S}\left(\mathbf{G}\right)\right\}$$

and let $\mathcal{V}_N\left(\mathbf{G}\right)$ be the linear subspace of $\mathbb{R}^{n^2N}$ defined by

$$\mathcal{V}_N\left(\mathbf{G}\right)=\left\{\mathbf{1}_N\otimes vec\left(\mathbf{G\Psi G}^{\top}\right):\mathbf{\Psi}\in\mathcal{S}_{|m|}\right\}=\left\{\mathbf{1}_N\otimes vec\left(\mathbf{\Gamma}\right):\mathbf{\Gamma}\in\mathcal{S}\left(\mathbf{G}\right)\right\},$$

where $\mathbf{1}_N=\left(1,...,1\right)^{\top}\in\mathbb{R}^N$, the symbol $\otimes$ denotes the Kronecker product and $vec\left(\mathbf{A}\right)$ is the vectorization of a matrix $\mathbf{A}$ (see their definitions in the Appendix). All these spaces are regarded as Euclidean spaces with the scalar product associated to the Frobenius matrix norm.

3.2.2 Model selection approach for covariance estimation

The approach that we will develop to estimate the covariance function σ is based on the following two main ingredients: first, we consider a functional expansion $\widetilde{X}$ to approximate the underlying process X and take the covariance of $\widetilde{X}$ as an approximation of the true covariance σ.

For this, let $m \in \mathcal{M}$ and consider an approximation to the process X of the following form:

$$\widetilde{X}\left(t\right) = \sum_{\lambda \in m} a_{\lambda} g_{\lambda}\left(t\right), \tag{3.5}$$

where a_{λ} are suitable random coefficients. For instance if X takes its values in $L^2(T)$ (the space of square integrable real-valued functions on T) and if $(g_{\lambda})_{\lambda \in m}$ are orthonormal functions in $L^2(T)$, then one can take

$$a_{\lambda} = \int_T X(t) g_{\lambda}(t) dt.$$

Several basis can thus be considered, such as a polynomial basis on $\mathbb{R}^d$, Fourier expansion on a rectangle $T \subset \mathbb{R}^d$ (i.e. $g_{\lambda}\left(t\right) = e^{i2\pi\langle\omega_{\lambda},t\rangle}$, using a regular grid of discrete set of frequencies $\left\{\omega_{\lambda} \in \mathbb{R}^d, \lambda \in m\right\}$ that do not depend on $t_1, ..., t_n$). One can also use, as in Elogne, Perrin and Thomas-Agnan [39], tensorial product of B-splines on a rectangle $T \subset \mathbb{R}^d$, with a regular grid of nodes in $\mathbb{R}^d$ not depending on $t_1, ..., t_n$ or a standard wavelet basis on $\mathbb{R}^d$, depending on a regular grid of locations in $\mathbb{R}^d$ and discrete scales in $\mathbb{R}_+$. Another class of natural expansion is provided by Karhunen-Loève expansion of the process X (see Adler [1] for more references).

Therefore, it is natural to consider the covariance function ρ of $\widetilde{X}$ as an approximation of σ. Since the covariance ρ can be written as

$$\rho\left(s,t\right) = \mathbf{G}_s^{\top} \overline{\mathbf{\Psi}} \mathbf{G}_t, \tag{3.6}$$

where, after reindexing the functions if necessary, $\mathbf{G}_t = \left(g_{\lambda}\left(t\right), \lambda \in m\right)^{\top}$ and

$$\overline{\mathbf{\Psi}} = \left(\mathbb{E}\left(a_{\lambda} a_{\mu}\right)\right), \text{ with } (\lambda, \mu) \in m \times m.$$

Hence we are led to look for an estimate $\widehat{\sigma}$ of σ in the class of functions of the form (3.6), with $\mathbf{\Psi} \in \mathbb{R}^{|m| \times |m|}$ some symmetric matrix. Note that the choice of the function expansion in (3.5), in particular the choice of the subset of indices m, will be crucial in the approximation properties of the covariance function ρ. This estimation procedure has several advantages: it will be shown that an appropriate choice of loss function leads to the construction of symmetric n.n.d. matrix $\widehat{\mathbf{\Psi}}$ (see Proposition 3.1) and thus the resulting estimate

$$\widehat{\sigma}\left(s,t\right) = \mathbf{G}_s^{\top} \widehat{\mathbf{\Psi}} \mathbf{G}_t$$

is a covariance function, so it can be plugged in other procedures which requires working with a covariance function. We also point out that the large amount of existing approaches for function approximation of the type (3.5) (such as those based on Fourier, wavelets, kernel, splines or radial functions) provides great flexibility to the model (3.6).

Secondly, we use the Frobenius matrix norm to quantify the risk of the covariance matrix estimators. Recall that $\mathbf{\Sigma} = \left(\sigma\left(t_j, t_k\right)\right)_{1\leq j,k\leq n}$ is the true covariance matrix while $\mathbf{\Gamma} = \left(\rho\left(t_j, t_k\right)\right)_{1\leq j,k\leq n}$ will denote the covariance matrix of the approximated process $\widetilde{X}$ at the observation points. Hence

$$\mathbf{\Gamma} = \mathbf{G}\overline{\mathbf{\Psi}}\mathbf{G}^{\top}. \tag{3.7}$$

Comparing the covariance function ρ with the true one σ over the design points t_j, implies quantifying the deviation of $\mathbf{\Gamma}$ from $\mathbf{\Sigma}$. For this consider the following loss function

$$L\left(\mathbf{\Psi}\right) = \mathbb{E}\left\|\mathbf{x}\mathbf{x}^{\top} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^{\top}\right\|_F^2,$$

where $\mathbf{x} = \left(X\left(t_1\right), ..., X\left(t_n\right)\right)^{\top}$ and $\|.\|_F$ is the Frobenius matrix norm. Note that

$$L\left(\mathbf{\Psi}\right) = \left\|\mathbf{\Sigma} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^{\top}\right\|_F^2 + C,$$

where the constant C does not depend on $\mathbf{\Psi}$. The Frobenius matrix norm provides a meaningful metric for comparing covariance matrices, widely used in multivariate analysis, in particular in the theory on principal components analysis. See also Biscay et al. [18], Schafer and Strimmer [81] and references therein for other applications of this loss function.

To the loss L corresponds the following empirical contrast function L_N, which will be the fitting criterion we will try to minimize

$$L_N\left(\mathbf{\Psi}\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i\mathbf{x}_i^{\top} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^{\top}\right\|_F^2.$$

We point out that this loss is exactly the sum of the squares of the residuals corresponding to the matrix linear regression model

$$\mathbf{x}_i\mathbf{x}_i^{\top} = \mathbf{G}\mathbf{\Psi}\mathbf{G}^{\top} + \mathbf{U}_i, \quad i = 1, ..., N, \tag{3.8}$$

with i.i.d. matrix errors $\mathbf{U}_i$ such that $\mathbb{E}\left(\mathbf{U}_i\right) = \mathbf{0}$. This remark provides a natural framework to study the covariance estimation problem as a matrix regression model. Note also that the set of matrices $\mathbf{G}\mathbf{\Psi}\mathbf{G}^{\top}$ is a linear subspace of $\mathbb{R}^{n\times n}$ when $\mathbf{\Psi}$ ranges over the space of symmetric matrices $\mathcal{S}_{|m|}$.

To summarize our approach, we finally propose following two-step estimation procedure: in a first step, for a given design matrix $\mathbf{G}$, define

$$\widehat{\mathbf{\Psi}} = \underset{\mathbf{\Psi}\in\mathcal{S}_{|m|}}{\arg\min}\, L_N(\mathbf{\Psi}),$$

and take $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^{\top}$ as an estimator of $\mathbf{\Sigma}$. Note that $\widehat{\mathbf{\Psi}}$ will be shown to be a n.n.d. matrix (see Proposition 3.1) and thus $\widehat{\mathbf{\Sigma}}$ is also a n.n.d. matrix. Since the minimization of $L_N\left(\mathbf{\Psi}\right)$ with respect to $\mathbf{\Psi}$ is done over the linear space of symmetric matrices $\mathcal{S}_{|m|}$, it can be transformed to a classical least squares linear problem, and the computation of $\widehat{\mathbf{\Psi}}$ is therefore quite simple. For a given design matrix $\mathbf{G}$, we will construct an estimator for $\mathbf{\Gamma} = \mathbf{G}\overline{\mathbf{\Psi}}\mathbf{G}^{\top}$ which will be close to $\mathbf{\Sigma} = \mathbb{V}\left(\mathbf{x}_1\right)$ as soon as $\widetilde{X}$ is a sharp approximation of X. So, the role of $\mathbf{G}$ and thus the choice of the subset of indices m is crucial since it determines the behavior of the estimator.

Hence, in a second step, we aim at selecting the best design matrix $\mathbf{G} = \mathbf{G}_m$ among a collection of candidates $\{\mathbf{G}_m, m \in \mathcal{M}\}$. For this, methods and results from the theory of model selection in linear regression can be applied to the present context. In particular the results in Baraud [7], Comte [27] or Loubes and Ludena [58] will be useful in dealing with model selection for the framework (3.8). Note that only assumptions about moments, not specific distributions of the data, are involved in the estimation procedure.

Remark 3.1. *We consider here a least squares estimates of the covariance. Note that suitable regularization terms or constraints could also be incorporated into the minimization of $L_N(\mathbf{\Psi})$ to impose desired properties for the resulting estimator, such as smoothness or sparsity conditions as in Levina, Rothman and Zhu [57].*

3.3 Oracle inequality for covariance estimation

The first part of this section describes the properties of the least squares estimator $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$, while the second part builds a selection procedure to pick automatically the best estimate among a collection of candidates.

3.3.1 Least squares covariance estimation

Given some $n \times |m|$ fixed design matrix $\mathbf{G}$ associated to a finite family of $|m|$ basis functions, the least squares covariance estimator of $\mathbf{\Sigma}$ is defined by

$$\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top = \underset{\mathbf{\Psi}\in\mathcal{S}_{|m|}}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2\right\}. \tag{3.9}$$

The corresponding estimator of the covariance function σ is

$$\widehat{\sigma}(s,t) = \mathbf{G}_s^\top\widehat{\mathbf{\Psi}}\mathbf{G}_t. \tag{3.10}$$

Proposition 3.1. *Let $\mathbf{Y}_1, ..., \mathbf{Y}_N \in \mathbb{R}^{n\times n}$ and $\mathbf{G} \in\mathbb{R}^{n\times|m|}$ be arbitrary matrices Then,*
(a) The infimum

$$\inf_{\mathbf{\Psi}\in\mathcal{S}_{|m|}}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{Y}_i - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2\right\}$$

is achieved at

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^\top\mathbf{G}\right)^{-}\mathbf{G}^\top\left(\frac{\overline{\mathbf{Y}} + \overline{\mathbf{Y}}^\top}{2}\right)\mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^{-}, \tag{3.11}$$

where $\left(\mathbf{G}^\top\mathbf{G}\right)^{-}$ is any generalized inverse of $\mathbf{G}^\top\mathbf{G}$ (see Seber [84] for a general definition), and

$$\overline{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i.$$

(b) Furthermore, $\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^{\top}$ *is the same for all the generalized inverses* $\left(\mathbf{G}^{\top}\mathbf{G}\right)^{-}$ *of* $\mathbf{G}^{\top}\mathbf{G}$*. In particular, if* $\mathbf{Y}_1, ..., \mathbf{Y}_N \in \mathcal{S}_n$ *(i.e., if they are symmetric matrices) then any minimizer has the form*

$$\widehat{\mathbf{\Psi}} = \left(\mathbf{G}^{\top}\mathbf{G}\right)^{-}\mathbf{G}^{\top}\overline{\mathbf{Y}}\mathbf{G}\left(\mathbf{G}^{\top}\mathbf{G}\right)^{-}.$$

If $\mathbf{Y}_1, ..., \mathbf{Y}_N$ *are n.n.d. then these matrices* $\widehat{\mathbf{\Psi}}$ *are n.n.d.*

If we assume that $(\mathbf{G}^{\top}\mathbf{G})^{-1}$ exists, then Proposition 3.1 shows that we retrieve the expression (3.3) for $\widehat{\mathbf{\Psi}}$ that has been derived from least squares estimation in model (3.1).

Theorem 3.1. *Let* $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\top}$*. Then, the least squares covariance estimator defined by (3.9) is given by the n.n.d. matrix*

$$\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^{\top} = \mathbf{\Pi}\mathbf{S}\mathbf{\Pi},$$

where

$$\begin{aligned}\widehat{\mathbf{\Psi}} &= \left(\mathbf{G}^{\top}\mathbf{G}\right)^{-}\mathbf{G}^{\top}\mathbf{S}\mathbf{G}\left(\mathbf{G}^{\top}\mathbf{G}\right)^{-} \\ \mathbf{\Pi} &= \mathbf{G}\left(\mathbf{G}^{\top}\mathbf{G}\right)^{-}\mathbf{G}^{\top}.\end{aligned} \tag{3.12}$$

Moreover $\widehat{\mathbf{\Sigma}}$ *has the following interpretations in terms of orthogonal projections:*

i) $\widehat{\mathbf{\Sigma}}$ *is the projection of* $\mathbf{S} \in \mathbb{R}^{n\times n}$ *on* $\mathcal{S}(\mathbf{G})$*.*

ii) $\mathbf{1}_N \otimes \widehat{\mathbf{\Sigma}}$ *is the projection of* $\mathbf{Y} = \left(\mathbf{x}_1\mathbf{x}_1^{\top}, ..., \mathbf{x}_N\mathbf{x}_N^{\top}\right)^{\top} \in \mathbb{R}^{nN\times n}$ *on* $\mathcal{S}_N(\mathbf{G})$*.*

iii) $\mathbf{1}_N \otimes vec\left(\widehat{\mathbf{\Sigma}}\right)$ *is the projection of* $\mathbf{y} = \left(vec^{\top}\left(\mathbf{x}_1\mathbf{x}_1^{\top}\right), ..., vec^{\top}\left(\mathbf{x}_N\mathbf{x}_N^{\top}\right)\right)^{\top} \in \mathbb{R}^{n^2N}$ *on* $\mathcal{V}_N(\mathbf{G})$*.*

The proof of this theorem is a direct application of Proposition 3.1. Hence for a given design matrix $\mathbf{G}$, the least squares estimator $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}(\mathbf{G})$ is well defined and has the structure of a covariance matrix. It remains to study how to pick automatically the estimate when dealing with a collection of design matrices coming from several approximation choices for the random process X.

3.3.2 Main result

Consider a collection of indices $m \in \mathcal{M}$ with size $|m|$. Let also $\{\mathbf{G}_m : m \in \mathcal{M}\}$ be a finite family of design matrices $\mathbf{G}_m \in \mathbb{R}^{n\times|m|}$, and let $\widehat{\mathbf{\Sigma}}_m = \widehat{\mathbf{\Sigma}}(\mathbf{G}_m)$, $m \in \mathcal{M}$, be the corresponding least squares covariance estimators. The problem of interest is to select the best of these estimators in the sense of the minimal quadratic risk $\mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2$.

The main theorem of this section provides a non-asymptotic bound for the risk of a penalized strategy for this problem. For all $m \in \mathcal{M}$, write

$$\begin{aligned}\mathbf{\Pi}_m &= \mathbf{G}_m\left(\mathbf{G}_m^{\top}\mathbf{G}_m\right)^{-}\mathbf{G}_m^{\top} \\ D_m &= \mathrm{Tr}\left(\mathbf{\Pi}_m\right),\end{aligned} \tag{3.13}$$

We assume that $D_m \geq 1$ for all $m \in \mathcal{M}$. The estimation error for a given model $m \in \mathcal{M}$ is given by

$$\mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2 = \left\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\right\|_F^2 + \frac{\delta_m^2 D_m}{N}, \tag{3.14}$$

where

$$\delta_m^2 = \frac{\mathrm{Tr}\left(\left(\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m\right)\mathbf{\Phi}\right)}{D_m}$$
$$\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right).$$

Given $\theta > 0$, define the penalized covariance estimator $\widetilde{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_{\widehat{m}}$ by

$$\widehat{m} = \arg\min_{m\in\mathcal{M}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \left\| \mathbf{x}_i\mathbf{x}_i^\top - \widehat{\mathbf{\Sigma}}_m \right\|_F^2 + pen\left(m\right) \right\},$$

where

$$pen\left(m\right) = \left(1+\theta\right)\frac{\delta_m^2 D_m}{N}. \tag{3.15}$$

Theorem 3.2. *Let $q > 0$ be a given constant such that there exists $p > 2\left(1+q\right)$ satisfying $\mathbb{E}\left\|\mathbf{x}_1\mathbf{x}_1^\top\right\|_F^p < \infty$. Then, for some constants $K\left(\theta\right) > 1$ and $C'\left(\theta, p, q\right) > 0$ we have that*

$$\left(\mathbb{E}\left\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\right\|_F^{2q}\right)^{1/q} \leq 2^{\left(q^{-1}-1\right)_+}\left[K\left(\theta\right)\inf_{m\in\mathcal{M}}\left(\left\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\right\|_F^2 + \frac{\delta_m^2 D_m}{N}\right) + \frac{\Delta_p}{N}\delta_{\sup}^2\right],$$

where

$$\Delta_p^q = C'\left(\theta, p, q\right)\mathbb{E}\left\|\mathbf{x}_1\mathbf{x}_1^\top\right\|_F^p\left(\sum_{m\in\mathcal{M}}\delta_m^{-p}D_m^{-\left(p/2-1-q\right)}\right)$$

and

$$\delta_{\sup}^2 = \max\left\{\delta_m^2 : m\in\mathcal{M}\right\}.$$

In particular, for $q = 1$ we have

$$\mathbb{E}\left(\left\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\right\|_F^2\right) \leq K\left(\theta\right)\inf_{m\in\mathcal{M}}\mathbb{E}\left(\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2\right) + \frac{\Delta_p}{N}\delta_{\sup}^2. \tag{3.16}$$

For the proof of this result, we first restate this theorem in a vectorized form which turns to be a k-variate extensions of results in Baraud [7] (which are covered when $k = 1$) and are stated in Section 3.4.1. Their proof rely on model selection techniques and a concentration tool stated in Section 3.4.2.

Remark 3.2. *Note that the penalty depends on the quantity δ_m which is unknown in practice. Indeed, the penalty relies on $\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$, which reflects the correlation structure of the data. In the original paper by Baraud [8], an estimator of the variance is proposed to overcome this issue. However, the consistency proof relies on a concentration inequality which turns to be a χ^2 like inequality. Extending this inequality to our case would mean to be able to construct concentration bounds for matrices $\mathbf{x}\mathbf{x}^\top$, implying Wishart distributions. Some results exist in this framework, see for instance Rao, Mingo, Speicher and Edelman [77], but adapting this kind of construction to our case is a hard task which falls beyond the scope of this work.*
However, we point out that for practical purpose, when N is large enough, this quantity

can be consistently estimated using the empirical version of $\mathbf{\Phi}$ *since the* $\mathbf{x}_i$*,* $i = 1, \ldots, N$ *are i.i.d. observed random variables, which is given by*

$$\widehat{\mathbf{\Phi}} = \frac{1}{N} \sum_{i=1}^{N} vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right) \left(vec\left(\mathbf{x}_i \mathbf{x}_i^\top\right)\right)^\top - vec(\mathbf{S}) \left(vec(\mathbf{S})\right)^\top . \tag{3.17}$$

Hence, there is a practical way of computing the penalty. The influence of the use of such an estimated penalty is studied in Section 3.5. Note also that if $\kappa > 0$ *denotes any bound of* δ_m^2 *such that* $\delta_m^2 \leq \kappa$ *for all* m*, then Theorem 3.2 remains true if* δ_m^2 *is replaced by* κ *in all the statements.*

We have obtained in Theorem 3.2 an oracle inequality since, using (3.14) and (3.16), one immediately sees that $\widetilde{\mathbf{\Sigma}}$ has the same quadratic risk as the "oracle" estimator except for an additive term of order $O\left(\frac{1}{N}\right)$ and a constant factor. Hence, the selection procedure is optimal in the sense that it behaves as if the true model were at hand. To describe the result in terms of rate of convergence, we have to pay a special attention to the bias terms $\|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|_F^2$. In a very general framework, it is difficult to evaluate such approximation terms. If the process has bounded second moments, i.e. for all $j = 1, \ldots, n$, we have $\mathbb{E}\left(X^2\left(t_j\right)\right) \leq C$, then we can write

$$\begin{aligned} \|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|_F^2 &\leq C \sum_{j=1}^{n} \sum_{j'=1}^{n} \left[\mathbb{E}\left(X\left(t_j\right) - \widetilde{X}\left(t_j\right)\right)^2 + \mathbb{E}\left(X\left(t_{j'}\right) - \widetilde{X}\left(t_{j'}\right)\right)^2\right] \\ &\leq 2Cn^2 \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left(X\left(t_j\right) - \widetilde{X}\left(t_j\right)\right)^2 . \end{aligned}$$

Since n is fixed and the asymptotics are given with respect to N, the number of replications of the process, the rate of convergence relies on the quadratic error of the expansion of the process. To compute the rate of convergence, this approximation error must be controlled.

From a theoretical point of view, take $d = 1$, $T = [a, b]$, and consider a process $X(t)$ with $t \in [a, b]$, for which the basis of its Karhunen-Loève expansion is known. Set $\mathcal{M} = \mathcal{M}_N = \{m = \{1, \ldots, |m|\}, |m| = 1, \ldots, N\}$. Then we can write $X(t) = \sum\limits_{\lambda=1}^{\infty} Z_\lambda g_\lambda(t)$, where Z_λ are centered random variables such that $\mathbb{E}\left(Z_\lambda^2\right) = \gamma_\lambda^2$, where γ_λ^2 is the eigenvalue corresponding to the eigenfunction g_λ of the operator $(Kf)(t) = \int\limits_a^b \sigma(s, t) f(s)\, ds$. If $X(t)$ is a Gaussian process then the random variables Z_λ are Gaussian and stochastically independent. Hence, a natural approximation of $X(t)$ is given by $\widetilde{X}(t) = \sum\limits_{\lambda=1}^{|m|} Z_\lambda g_\lambda(t)$. So we have that

$$\mathbb{E}\left(X(t) - \widetilde{X}(t)\right)^2 = \mathbb{E}\left(\sum_{\lambda=|m|+1}^{\infty} Z_\lambda g_\lambda(t)\right)^2 = \sum_{\lambda=|m|+1}^{\infty} \gamma_\lambda^2 g_\lambda^2(t) .$$

Therefore, if $\|g_\lambda\|^2_{L_2([a,b])} = 1$ then $\mathbb{E}\left\|X(t) - \widetilde{X}(t)\right\|^2_{L_2([a,b])} = \sum\limits_{\lambda=|m|+1}^{\infty} \gamma^2_\lambda$. Assume that the γ_λ's have a polynomial decay of rate $\alpha > 0$, namely $\gamma_\lambda \sim \lambda^{-\alpha}$, then we get an approximation error of order $O\left((|m|+1)^{-2\alpha}\right)$. Hence, we get that (under appropriate conditions on the design points $t_1, \ldots, t_n$)

$$\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\|^2_F = O\left((|m|+1)^{-2\alpha}\right).$$

Finally, since in this example

$$\mathbb{E}\left\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\right\|^2_F \leq K(\theta) \inf_{m\in\mathcal{M}_N} \left(\|\mathbf{\Sigma} - \mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\|^2_F + \frac{\delta^2_m m}{N}\right) + O\left(\frac{1}{N}\right)$$

then the quadratic risk is of order $N^{-\frac{2\alpha}{2\alpha+1}}$ as soon as $|m| \sim N^{1/(2\alpha+1)}$ belongs to the collection of models $\mathcal{M}_N$. In another framework, if we consider a spline expansion, the rate of convergence for the approximation given in Elogne, Perrin and Thomas-Agnan [39] are of the same order.

Hence we have obtained a model selection procedure which enables to recover the best covariance model among a given collection. This method works without strong assumptions on the process, in particular stationarity is not assumed, but at the expend of necessary i.i.d. observations of the process at the same points.
We point out that this study requires a large number of replications N with respect to the number of observation points n. Moreover, since for a practical use of this methodology, an estimator of the penalty must be computed, relying on the estimation of the 4-th order moment, the need for a large amount of data is crucial even if the simulations are still, quite satisfactory, for not so large sample. This setting is quite common in epidemiology where a phenomenon is studied at a large number of locations but only during a short time. Hence our method is not designed to tackle the problem of covariance estimation in the high dimensional case $n >> N$. This topic has received a growing attention over the past years and we refer to Bickel and Levina [12] and references therein for a survey.

3.4 Model selection for multidimensional regression

3.4.1 Oracle inequality for multidimensional regression model

Recall that we consider the following model

$$\mathbf{x}_i\mathbf{x}_i^\top = \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top + \mathbf{U}_i, \quad i = 1, ..., N,$$

with i.i.d. matrix errors $\mathbf{U}_i \in \mathbb{R}^{n\times n}$ such that $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$.

The key point is that previous model can be rewritten in vectorized form in the following way

$$\mathbf{y}_i = \mathbf{A}\beta + \mathbf{u}_i, \quad i = 1, ..., N, \tag{3.18}$$

where $\mathbf{y}_i = vec\left(\mathbf{x}_i\mathbf{x}_i^\top\right) \in \mathbb{R}^{n^2}$, $\mathbf{A} = (\mathbf{G}\otimes\mathbf{G})\,\mathbf{D}_m \in \mathbb{R}^{n^2\times\frac{|m|(|m|+1)}{2}}$, where $\mathbf{D}_m \in \mathbb{R}^{|m|^2\times\frac{|m|(|m|+1)}{2}}$ is the duplication matrix, $\beta = vech(\mathbf{\Psi}) \in \mathbb{R}^{\frac{|m|(|m|+1)}{2}}$, and $\mathbf{u}_i = vec(\mathbf{U}_i) \in \mathbb{R}^{n^2}$(see definitions A.4, A.6, A.5 and property (A.54) in the Appendix).

Note that this model is equivalent to the following regression model

$$\mathbf{y} = \left(\mathbf{1}_N \otimes \mathbf{A}\right)\beta + \mathbf{u}, \tag{3.19}$$

where $\mathbf{y} = \left(\left(\mathbf{y}_1\right)^\top, ..., \left(\mathbf{y}_N\right)^\top\right)^\top \in \mathbb{R}^{Nn^2}$ is the data vector, $\left(\mathbf{1}_N \otimes \mathbf{A}\right) \in \mathbb{R}^{Nn^2 \times \frac{|m|(|m|+1)}{2}}$ is a known fixed matrix, $\beta = vech\left(\mathbf{\Psi}\right)$ is an unknown vector parameter as before, and $\mathbf{u} = \left(\left(\mathbf{u}_1\right)^\top, ..., \left(\mathbf{u}_N\right)^\top\right)^\top \in \mathbb{R}^{Nn^2}$ is such that $\mathbb{E}\left(\mathbf{u}\right) = \mathbf{0}$.

It is worth of noting that this regression model has several peculiarities in comparison with standard ones.
i) The error $\mathbf{u}$ has a specific correlation structure, namely $\mathbf{I}_N \otimes \mathbf{\Phi}$, where $\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$.
ii) In contrast with standard multivariate models, each coordinate of $\mathbf{y}$ depends on all the coordinates of β.
iii) For any estimator $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ that be a linear function of the sample covariance $\mathbf{S}$ of the data $\mathbf{x}_1$,...,$\mathbf{x}_N$ (and so, in particular, for the estimator minimizing L_N) it is possible to construct an unbiased estimator of its quadratic risk $\mathbb{E}\left\|\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}\right\|_F^2$.

More generally, assume we observe $\mathbf{y}_i$, $i = 1, \ldots, N$ random vectors of $\mathbb{R}^k$, with $k \geq 1$ ($k = n^2$ in the particular case of model (3.18)), such that

$$\mathbf{y}_i = \mathbf{f}^i + \boldsymbol{\varepsilon}_i, \quad i = 1, ..., N, \tag{3.20}$$

where $\mathbf{f}^i \in \mathbb{R}^k$ are nonrandom and $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_N$ are i.i.d. random vectors in $\mathbb{R}^k$ with $\mathbb{E}\left(\boldsymbol{\varepsilon}_1\right) = \mathbf{0}$ and $\mathbb{V}\left(\boldsymbol{\varepsilon}_1\right) = \mathbf{\Phi}$. For sake of simplicity, we identify the function $g : \mathcal{X} \to \mathbb{R}^k$ with vectors $\left(g\left(x_1\right), \ldots, g\left(x_N\right)\right)^\top \in \mathbb{R}^{Nk}$ and we denote by $\left\langle \mathbf{a}, \mathbf{b}\right\rangle_N = \frac{1}{N}\sum_{i=1}^N \mathbf{a}_i^\top \mathbf{b}_i$ the inner product of $\mathbb{R}^{Nk}$ associated to the norm $\|.\|_N$ defined by $\|\mathbf{a}\|_N^2 = \frac{1}{N}\sum_{i=1}^N \mathbf{a}_i^\top \mathbf{a}_i = \frac{1}{N}\sum_{i=1}^N \|\mathbf{a}_i\|_{\ell_2}^2$, where $\mathbf{a} = \left(\mathbf{a}_1^\top, \ldots, \mathbf{a}_N^\top\right)^\top$ and $\mathbf{b} = \left(\mathbf{b}_1^\top, \ldots, \mathbf{b}_N^\top\right)^\top$ with $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^k$ for all $i = 1, ..., N$.

Given $N, k \in \mathbb{N}$, let $\left(\mathcal{L}_m\right)_{m \in \mathcal{M}}$ be a finite family of linear subspaces of $\mathbb{R}^{Nk}$. For each $m \in \mathcal{M}$, assume $\mathcal{L}_m$ has dimension $D_m \geq 1$. Let $\widehat{\mathbf{f}}_m$ be the least squares estimator of $\mathbf{f} = \left(\left(\mathbf{f}^1\right)^\top, ..., \left(\mathbf{f}^N\right)^\top\right)^\top$ based on the data $\mathbf{y} = \left(\mathbf{y}_1^\top, ..., \mathbf{y}_N^\top\right)^\top$ under the model $\mathcal{L}_m$, i.e.

$$\widehat{\mathbf{f}}_m = \arg\min_{\mathbf{v} \in \mathcal{L}_m} \|\mathbf{y} - \mathbf{v}\|_N^2 = \mathbf{P}_m \mathbf{y},$$

where $\mathbf{P}_m$ is the orthogonal projection matrix from $\mathbb{R}^{Nk}$ on $\mathcal{L}_m$. Write

$$\delta_m^2 = \frac{\mathrm{Tr}\left(\mathbf{P}_m\left(\mathbf{I}_N \otimes \mathbf{\Phi}\right)\right)}{D_m}$$
$$\delta_{\sup}^2 = \max\left\{\delta_m^2 : m \in \mathcal{M}\right\}.$$

Given $\theta > 0$, define the penalized estimator $\widetilde{\mathbf{f}} = \widehat{\mathbf{f}}_{\widehat{m}}$, where

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} \left\{\left\|\mathbf{y}-\widehat{\mathbf{f}}_m\right\|_N^2 + pen\left(m\right)\right\},$$

with $pen\left(m\right) = \left(1 + \theta\right)\frac{\delta_m^2 D_m}{N}$.

Proposition 3.2. *Let $q > 0$ be given such that there exists $p > 2(1+q)$ satisfying $\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p < \infty$. Then, for some constants $K(\theta) > 1$ and $C(\theta, p, q) > 0$ we have that*

$$\mathbb{E}\left(\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - K(\theta)\, M_N^*\right)_+^q \leq \Delta_p^q \frac{\delta_{\sup}^{2q}}{N^q}, \tag{3.21}$$

where

$$\Delta_p^q = C(\theta, p, q)\, \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p \left(\sum_{m \in \mathcal{M}} \delta_m^{-p} D_m^{-(p/2-1-q)}\right)$$

$$M_N^* = \inf_{m \in \mathcal{M}} \left\{\left\|\mathbf{f}-\mathbf{P}_m \mathbf{f}\right\|_N^2 + \frac{\delta_m^2 D_m}{N}\right\}.$$

This theorem is equivalent to Theorem 3.2 using the vectorized version of the model (3.20) and turns to be an extension of Theorem 3.1 in Baraud [7] to the multivariate case. In a similar way, the following result constitutes also a natural extension of Corollary 3.1 in Baraud [7]. It is also closely related to the recent work in Gendre [44].

Corollary 3.1. *Under the assumptions of Proposition 3.2 it holds that*

$$\left(\mathbb{E}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^{2q}\right)^{1/q} \leq 2^{\left(q^{-1}-1\right)_+}\left[K(\theta) \inf_{m \in \mathcal{M}}\left(\left\|\mathbf{f}-\mathbf{P}_m \mathbf{f}\right\|_N^2 + \frac{\delta_m^2 D_m}{N}\right) + \frac{\Delta_p}{N} \delta_{\sup}^2\right],$$

where Δ_p was defined in Proposition 3.2.

Under regularity assumptions for the function $\mathbf{f}$, depending on a smoothness parameter s, the bias term is of order $\left\|\mathbf{f}-\mathbf{P}_m \mathbf{f}\right\|_N^2 = O(D_m^{-2s})$. Hence, for $q = 1$ we obtain the usual rate of convergence $N^{-\frac{2s}{2s+1}}$ for the quadratic risk as soon as the optimal choice $D_m = N^{\frac{1}{2s+1}}$ belongs to the collection of models, yielding the optimal rate of convergence for the penalized estimator.

3.4.2 Concentration bound for random processes

Recall that $k \geq 1$. The following result is a k-variate extension of results in Baraud [7] (which are covered when $k = 1$). Its proof is deferred to the end of the chapter.

Proposition 3.3. *(Extension of Corollary 5.1 in Baraud [7]). Given $N, k \in \mathbb{N}$, let the matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{Nk \times Nk} \backslash \{\mathbf{0}\}$ be symmetric and non-negative definite, and let $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_N$ be i.i.d. random vectors in $\mathbb{R}^k$ with $\mathbb{E}(\boldsymbol{\varepsilon}_1) = 0$ and $\mathbb{V}(\boldsymbol{\varepsilon}_1) = \boldsymbol{\Phi}$. Write $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_1^\top, ..., \boldsymbol{\varepsilon}_N^\top\right)^\top$, $\zeta(\boldsymbol{\varepsilon}) = \sqrt{\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{A}} \boldsymbol{\varepsilon}}$, and $\delta^2 = \frac{\operatorname{Tr}\left(\widetilde{\mathbf{A}}(\mathbf{I}_N \otimes \boldsymbol{\Phi})\right)}{\operatorname{Tr}\left(\widetilde{\mathbf{A}}\right)}$. For all $p \geq 2$ such that $\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p < \infty$ it holds that, for all $x > 0$,*

$$\mathbb{P}\left(\zeta^2(\boldsymbol{\varepsilon}) \geq \delta^2 \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) + 2\delta^2 \sqrt{\operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \tau\left(\widetilde{\mathbf{A}}\right) x} + \delta^2 \tau\left(\widetilde{\mathbf{A}}\right) x\right) \leq C(p) \frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right)}{\delta^p \tau\left(\widetilde{\mathbf{A}}\right) x^{p/2}}, \tag{3.22}$$

where the constant $C(p)$ depends only on p and $\tau\left(\widetilde{\mathbf{A}}\right)$ is the spectral radius of $\widetilde{\mathbf{A}}$.

Proposition 3.3 reduces to Corollary 5.1 in Baraud [7] when we only consider $k = 1$, in which case $\delta^2 = (\mathbf{\Phi})_{11} = \sigma^2$ is the variance of the univariate i.i.d. errors ε_i.

3.5 Numerical examples

In this section we illustrate the practical behaviour of the covariance estimator by model selection proposed in this chapter. In particular, we study its performance when computing the criterion using the estimated penalty described in Section 3.3.2. The programs for our simulations were implemented using MATLAB.

We will consider i.i.d. copies $X_1, \ldots, X_n$ of different Gaussian processes X on $T = [0,1]$ with values in $\mathbb{R}$, observed at fixed equi-spaced points $t_1, ..., t_n$ in $[0,1]$ for a fixed n, generated according to

$$X(t_j) = \sum_{\lambda=1}^{m^*} a_\lambda g_\lambda^* (t_j), \; j = 1, ..., n, \tag{3.23}$$

where m^* denotes the true model dimension, g_λ^* with $\lambda = 1, ..., m^*$ are orthonormal functions on $[0,1]$, and the coefficients $a_1, ..., a_{m^*}$ are independent and identically distributed Gaussian variables with zero mean. Note that $\mathbb{E}(X(t_j)) = 0$ for all $j = 1, ..., n$, the covariance function of the process X at the points $t_1, ..., t_n$ is given by

$$\sigma(t_j, t_k) = Cov(X(t_j), X(t_k)) = \sum_{\lambda=1}^{m^*} \mathbb{V}(a_\lambda) g_\lambda^*(t_j) g_\lambda^*(t_k)$$

for $1 \leq j,k \leq n$, with corresponding covariance matrix $\mathbf{\Sigma} = (\sigma(t_j, t_k))_{1 \leq j,k \leq n}$. Set $\mathbf{X} = (X_i(t_j))$ an $n \times N$ matrix, the columns of $\mathbf{X}$ are denoted by $\mathbf{x}_i = (X_i(t_1), ..., X_i(t_n))^\top$.

The covariance estimation by model selection is computed as follows. Let (g_λ) with $\lambda \in \mathbb{N}$ be an orthonormal basis on $[0,1]$ (wich may differ from the original basis functions g_λ^*, $\lambda = 1, ..., m^*$). For a given $M > 0$, candidate models are chosen among the collection $\mathcal{M} = \{\{1, ..., m\} : m = 1, ..., M\}$. To each set indexed by m we associate the matrix (model) $\mathbf{G}_m \in \mathbb{R}^{n \times m}$, with entries $(g_\lambda(t_j))_{1 \leq j \leq n, 1 \leq \lambda \leq m}$, which corresponds to a number m of basis functions $g_1, ..., g_m$ in the expansion to approximate the process X. We aim at choosing a good model among the family of models $\{\mathbf{G}_m : m \in \mathcal{M}\}$ in the sense of achieving the minimum of the quadratic risk

$$R(m) = \mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_m\right\|_F^2 = \|\mathbf{\Sigma} - \mathbf{\Pi}_m \mathbf{\Sigma} \mathbf{\Pi}_m\|_F^2 + \frac{\delta_m^2 D_m}{N}. \tag{3.24}$$

The ideal model m_0 is the minimizer of the risk function $m \mapsto R(m)$. Note that for all $m = 1, ..., M$,

$$L_N(m) = \frac{1}{N} \sum_{i=1}^{N} \left\|\mathbf{x}_i \mathbf{x}_i^\top - \widehat{\mathbf{\Sigma}}_m\right\|_F^2 = \mathcal{L}_N(m) + C$$

and

$$L_N(m) + pen(m) = PC(m) + C,$$

where

$$\begin{aligned} \mathcal{L}_N(m) &= \left\|\mathbf{S}-\widehat{\mathbf{\Sigma}}_m\right\|_F^2 = \|\mathbf{S}-\mathbf{\Pi}_m\mathbf{S}\mathbf{\Pi}_m\|_F^2 \\ PC(m) &= \left\|\mathbf{S}-\widehat{\mathbf{\Sigma}}_m\right\|_F^2 + pen(m) = \|\mathbf{S}-\mathbf{\Pi}_m\mathbf{S}\mathbf{\Pi}_m\|_F^2 + (1+\theta)\frac{\delta_m^2 D_m}{N}, \end{aligned} \tag{3.25}$$

the matrix $\widehat{\mathbf{\Sigma}}_m$ is the least squares covariance estimator of $\mathbf{\Sigma}$ corresponding to the model m as in Theorem 3.1 and the constant C does not depend on m. Thus, $\mathcal{L}_N$ and PC can be regarded as the empirical contrast function and the penalized criterion respectively that will be used for visual presentations of the results.

For each model $m = 1, ..., M$ we evaluate the penalized criterion (3.25) with $\theta = 1$ and expect that the minimum of PC is attained at a value $\widehat{m}(\delta^2)$ close to m_0. The quantity

$$\delta_m^2 = \frac{\mathrm{Tr}\left((\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m)\,\mathbf{\Phi}\right)}{D_m}$$

depends on the matrix $\mathbf{\Phi} = \mathbb{V}\left(vec\left(\mathbf{x}_1\mathbf{x}_1^\top\right)\right)$ as pointed out in Section 3.3.2), which is unknown in practice but can be consistently estimated by (3.17), yielding the plug-in estimate

$$\widehat{\delta}_m^2 = \frac{\mathrm{Tr}\left((\mathbf{\Pi}_m \otimes \mathbf{\Pi}_m)\,\widehat{\mathbf{\Phi}}\right)}{D_m}$$

for δ_m^2. We study the influence of using $\widehat{\delta}^2 = \left(\widehat{\delta}_m^2\right)_{1\leq m\leq M}$ rather than $\delta^2 = (\delta_m^2)_{1\leq m\leq M}$ on the model selection procedure.

Actually, we first compute the following approximation of the risk R,

$$\widehat{R}(m) = \|\mathbf{\Sigma}-\mathbf{\Pi}_m\mathbf{\Sigma}\mathbf{\Pi}_m\|_F^2 + \frac{\widehat{\delta}_m^2 D_m}{N},$$

and then, compute the estimator of the penalized criterion PC

$$\widehat{PC}(m) = \|\mathbf{S}-\mathbf{\Pi}_m\mathbf{S}\mathbf{\Pi}_m\|_F^2 + (1+\theta)\frac{\widehat{\delta}_m^2 D_m}{N}.$$

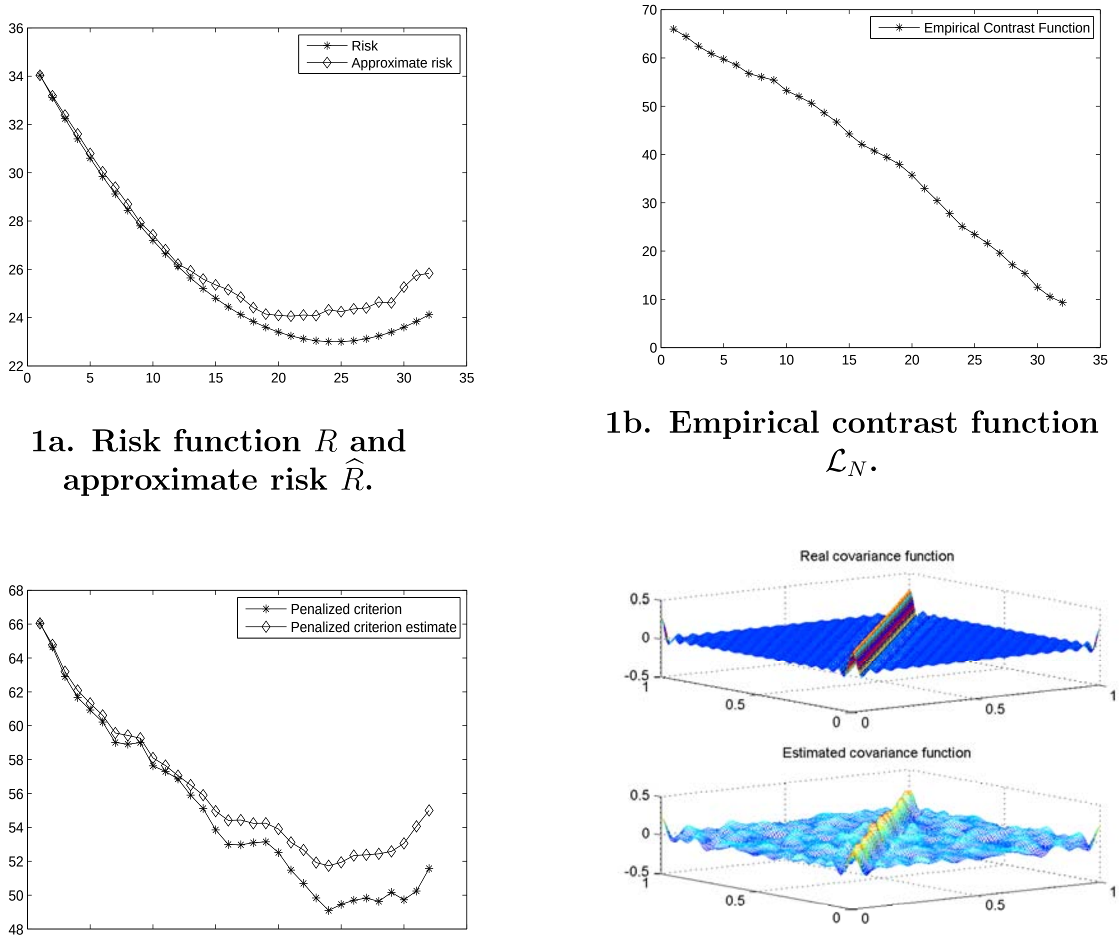
We denote by $\widehat{m}\left(\widehat{\delta}^2\right)$ the point at which the penalized criterion estimate $\widehat{PC}$ attains its minimum value, i.e., the model selected by minimizing $\widehat{PC}$.

In the following examples we plot the empirical contrast function $\mathcal{L}_N$ $(m = 1, ..., M)$, the risk function R, the approximate risk function $\widehat{R}$, the penalized criterion PC and the penalized criterion estimate $\widehat{PC}$. We also show figures of the true covariance function $\sigma(t,s)$ for $s,t \in [0,1]$ and the penalized covariance estimate based on $\widehat{PC}$, i.e., $\widehat{\sigma}(t,s) = \mathbf{G}_{\widehat{m},t}^\top \widehat{\mathbf{\Psi}}_{\widehat{m}} \mathbf{G}_{\widehat{m},s}$, where $\widehat{m} = \widehat{m}\left(\widehat{\delta}^2\right)$, $\widehat{\mathbf{\Psi}}_{\widehat{m}}$ is obtained as in Theorem 3.1 and $\mathbf{G}_{\widehat{m},t} = \left(g_1(t), ..., g_{\widehat{m}}(t)\right)^\top \in \mathbb{R}^{\widehat{m}}$ for all $t \in [0,1]$. Furthermore, we will focus attention on finite sample settings, i.e., those in which the number of repetitions N is not notably large (in comparison with the number n of design points t_j).

Example 1: Let $g_1^*, ..., g_{m^*}^*$ be the Fourier basis functions given by

$$g_\lambda^*(t) = \begin{cases} \frac{1}{\sqrt{n}} \text{ if } \lambda = 1 \\ \sqrt{2}\frac{1}{\sqrt{n}} \cos\left(2\pi\frac{\lambda}{2}t\right) \text{ if } \frac{\lambda}{2} \in \mathbb{Z} \\ \sqrt{2}\frac{1}{\sqrt{n}} \sin\left(2\pi\frac{\lambda-1}{2}t\right) \text{ if } \frac{\lambda-1}{2} \in \mathbb{Z}_* \end{cases} . \qquad (3.26)$$

We simulate a sample of size $N = 50$ according to (3.23) with $n = m^* = 35$ and $\mathbb{V}(a_\lambda) = 1$ for all $\lambda = 1, ..., m^*$. Set $M = 31$ and consider the models obtained by choosing m Fourier basis functions. In this setting, it can be shown that the minimum of the quadratic risk R is attained at $m_0 = \frac{N}{2} - 1$, which for $N = 50$ gives $m_0 = 24$. Figures 1a, 1b, 1c and 1d present the results obtained for a simulated sample.



1a. Risk function R and approximate risk $\widehat{R}$.

1b. Empirical contrast function $\mathcal{L}_N$.

1c. Penalized criterion PC and its estimate $\widehat{PC}$.

1d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

Figure 1a-1d: Results of Example 1 for a simulated sample.

Figure 1a shows that the approximate risk function $\widehat{R}$ reproduces the shape of the risk function R, so replacing δ^2 by $\widehat{\delta}^2$ into the risk does not have a too drastic effect. It can be observed in Figure 1b that, as expected, the empirical contrast function $\mathcal{L}_N$ is strictly decreasing over the whole range of possible models, hence its minimization would lead to choose the largest model $M = 31$. Interestingly, note that, unlike to what is quite common for the univariate linear regression with i.i.d. errors, the empirical contrast curve does not have an "elbow" (i.e., a change of curvature) around the optimal model $m_0 = 24$, which could provide by visual inspection some hint for selecting a suitable model.

On the contrary, both minimization of the penalized criterion PC and its estimate $\widehat{PC}$ lead to select the best model, i.e., $\widehat{m}\left(\delta^2\right) = 24$ and $\widehat{m}\left(\widehat{\delta}^2\right) = 24$ (see Figure 1c). This also demonstrates that replacing δ^2 by $\widehat{\delta}^2$ into the penalized criterion does not notably deteriorate the performance of the model selection procedure in this example.

Figure 1d shows that, in spite of the small sample size $N = 50$, a quite nice approximation to the true covariance function σ is achieved by its penalized covariance estimate $\widehat{\sigma}$ based on $\widehat{PC}$.

It is clear that the selected model $\widehat{m}\left(\widehat{\delta}^2\right)$ is a random variable that depends on the observed sample $\mathbf{X}$ through the penalized criterion estimate $\widehat{PC}$. Figure 1e illustrates such a variability by plotting the curves $\widehat{PC}$ corresponding to several simulated samples. It can be observed that the selected model $\widehat{m}\left(\widehat{\delta}^2\right)$ is close to the ideal model m_0, and the risk R evaluated at the selected model is much less than that of the largest model $M = 31$ that would be chosen by using the empirical contrast function.

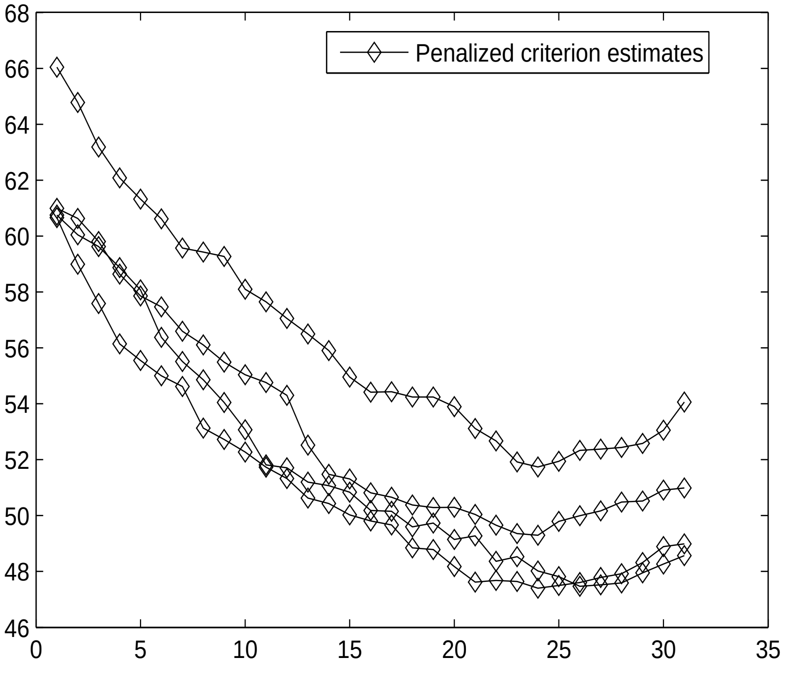
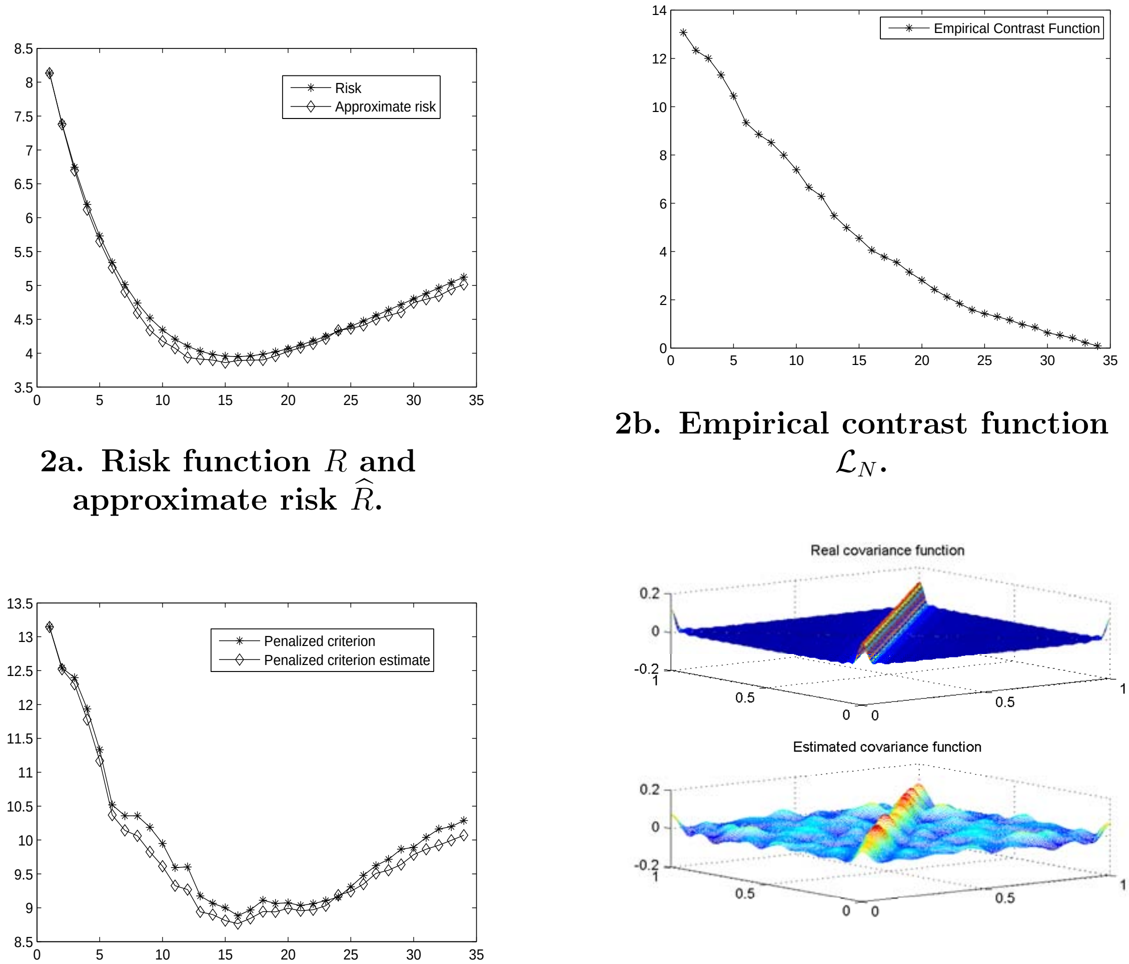


Figure 1e: Variability of the simulation process.

Example 2: Using the Fourier basis (3.26) we simulate a sample of size $N = 50$ according to (3.23) with $n = m^* = 35$ as in the previous example, but now we set a geometric decay of the variances $\mathbb{V}(a_\lambda)$, $\lambda = 1, ..., m^*$ (or equivalently, of the eigenvalues of the covariance operator of the process X); namely, $\mathbb{V}(a_1) = r$ and $\mathbb{V}(a_{\lambda+1}) = \mathbb{V}(a_\lambda)\, r$ for $\lambda = 2, 3, ...$, where $r = 0.95$. We consider a collection of models up to $M = 34$, with m Fourier basis functions. In this setting it can be proved that the minimum of the risk R is attained at $m_0 = (\log(2/[(1-r)(N-2)+2]))/\log(r)$, which yields $m_0 = 16$ for the actual values $N = 50$ and $r = 0.95$. The results obtained from a simulated sample are shown in Figures 2a, 2b, 2c and 2d. It can be noted that the empirical contrast function is strictly decreasing without any "elbow" effect, while the selected model by both the penalized criterion and the penalized criterion estimate is $\widehat{m}(\delta^2) = \widehat{m}\left(\widehat{\delta}^2\right) = 16$, which is the best model m_0 according to the risk R.



2a. Risk function R and approximate risk $\widehat{R}$.

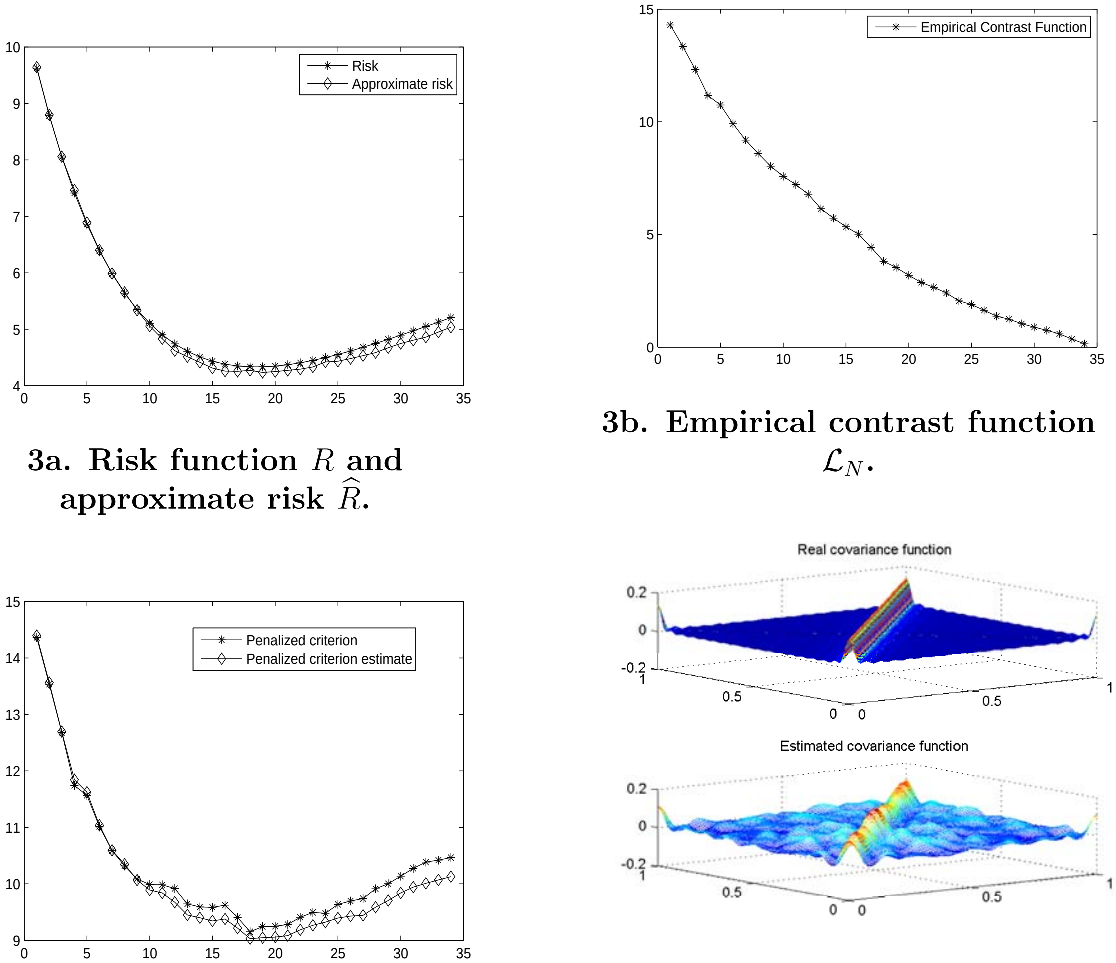
2b. Empirical contrast function $\mathcal{L}_N$.

2c. Penalized criterion PC and its estimate $\widehat{PC}$.

2d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

Figure 2: Results of Example 2.

Example 3: Using the Fourier basis (3.26) we simulate a sample of size $N = 60$ according to (3.23) with $n = m^* = 35$, but now we set the variances (eigenvalues) as follows: $\mathbb{V}\left(a_\lambda\right) = \sigma^2 + r^\lambda$ for all $\lambda = 1, ..., m^*$, where $r = 0.95$ and $\sigma^2 = 0.0475$. This decay of the eigenvalues is common in the *factor* models. Actually all the eigenvalues have almost the same small value σ^2 (corresponding to "noise") except for a few first eigenvalues that have larger values (corresponding to some "factors"). The collection of models considered corresponds to a number m $(1 \leq m \leq M)$ of Fourier basis functions up to $M = 34$. The results from a simulated sample are shown in Figures 3a, 3b, 3c and 3d. Figure 3a shows that the minimum of the risk function R is attained at $m_0 = 18$. Likewise the previous examples, the empirical contrast function is strictly decreasing without any "elbow", while the model selection procedure chooses the model $\widehat{m}\left(\delta^2\right) = \widehat{m}\left(\widehat{\delta^2}\right) = 18$, which is the value of m_0.



3a. Risk function R and approximate risk $\widehat{R}$.

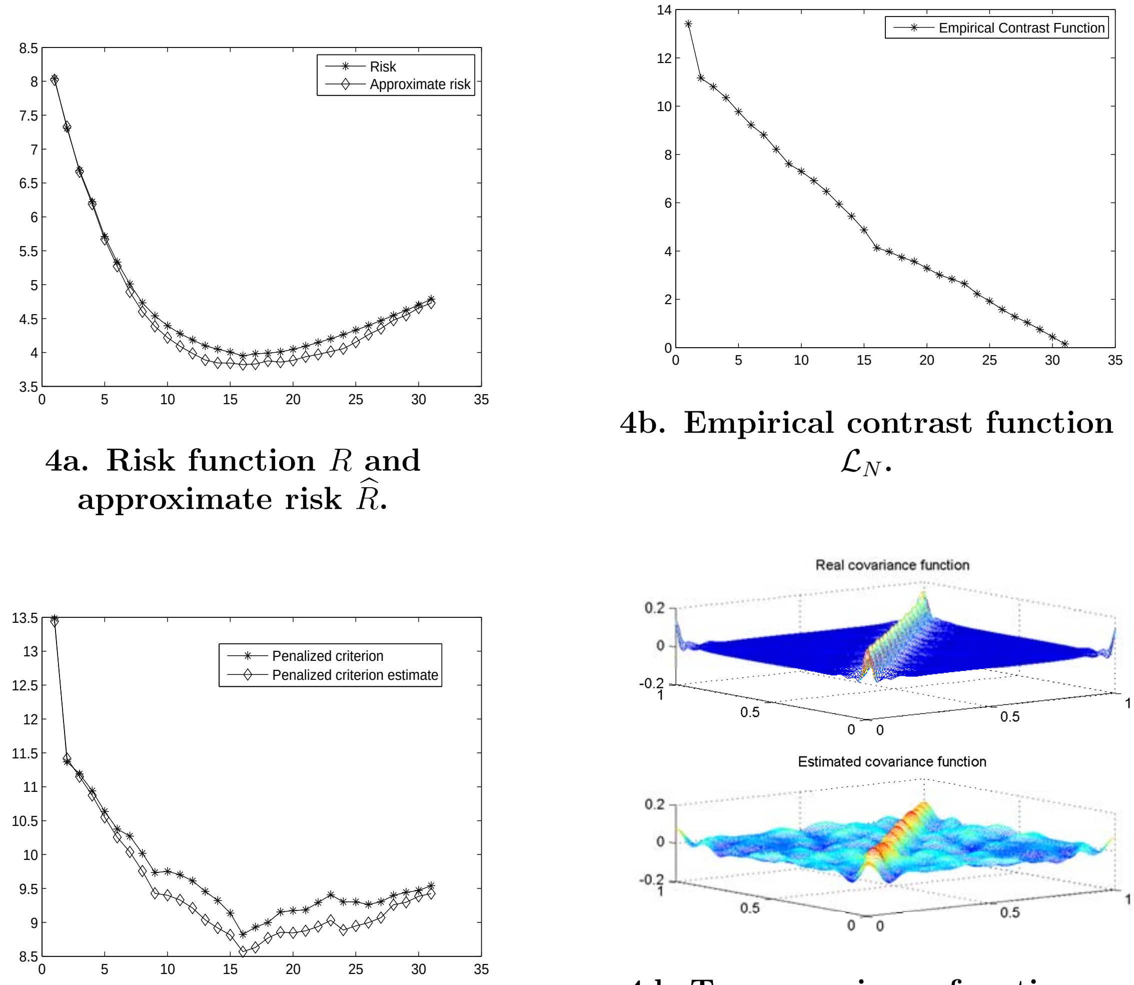
3b. Empirical contrast function $\mathcal{L}_N$.

3c. Penalized criterion PC and its estimate $\widehat{PC}$.

3d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

Figure 3: Results of Example 3.

Example 4: In this example we use different basis functions for generating the data and for estimating the covariance. Specifically, the process is generated using a wavelet basis and the collection of models considered in the model selection procedure corresponds to different numbers of Fourier basis functions up to $M = 31$. We simulate a sample of size $N = 50$ according to (3.23) using the Symmlet 8 wavelet basis, with $n = m^* = 32$. We set the variances of the random coefficients a_λ with a geometric decay likewise in Example 2, i.e., $\mathbb{V}\left(a_1\right) = r$ and $\mathbb{V}\left(a_{\lambda+1}\right) = \mathbb{V}\left(a_\lambda\right) r$, where $r = 0.95$. The results of one simulation are displayed in Figures 4a, 4b, 4c and 4d. Here it can be also observed that the penalized estimation procedure shows good performance even when using an estimate of the penalized criterion, leading to choosing the model $\widehat{m}\left(\widehat{\delta}^2\right) = \widehat{m}\left(\delta^2\right) = 16 = m_0$.



4a. Risk function R and approximate risk $\widehat{R}$.

4b. Empirical contrast function $\mathcal{L}_N$.

4c. Penalized criterion PC and its estimate $\widehat{PC}$.

4d. True covariance function σ and estimated covariance function $\widehat{\sigma}$.

Figure 4: Results of Example 4.

Summarizing the results of these simulated examples, we may conclude that for not so large sample sizes N:

$a)$ The empirical contrast function $\mathcal{L}_N$ is useless to select a model that attains a low risk. It is a strictly decreasing function whose minimization leads to simply choose the largest model M within the set of candidate models $m = 1, ..., M$. Furthermore, frequently the curve $\mathcal{L}_N$ does not have an "elbow" that could guide researchers to choose a suitable model by exploratory analysis.

$b)$ The covariance function estimator by model selection introduced in this chapter shows good performance in a variety of examples when based on the penalized criterion PC but also when using the estimated penalty $\widehat{PC}$.

3.6 Proofs

3.6.1 Proofs of preliminar results

Proof of Proposition 3.1.

$a)$ The minimization problem posed in this proposition is equivalent to minimize

$$h(\mathbf{\Psi}) = \left\|\overline{\mathbf{Y}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2.$$

The Frobenius norm $\|.\|_F$ is invariant by the vec operation. Furthermore, $\mathbf{\Psi} \in \mathcal{S}_{|m|}$ can be represented by means of $vec(\mathbf{\Psi}) = \mathbf{D}_m\beta$ where $\beta = vech(\mathbf{\Psi}) \in \mathbb{R}^{|m|(|m|+1)/2}$ and $\mathbf{D}_m$ is the duplication matrix. These facts and the property (A.54) in the Appendix allow one to rewrite

$$h(\mathbf{\Psi}) = \|\overline{\mathbf{y}} - (\mathbf{G}\otimes\mathbf{G})\mathbf{D}_m\beta\|_F^2 = \|\overline{\mathbf{y}} - (\mathbf{G}\otimes\mathbf{G})\mathbf{D}_m\beta\|_{\ell_2}^2,$$

where $\overline{\mathbf{y}} = vec\left(\overline{\mathbf{Y}}\right)$. Minimization of this quadratic function with respect to β in $\mathbb{R}^{|m|(|m|+1)/2}$ is equivalent to solve the normal equation

$$\mathbf{D}_m^\top(\mathbf{G}\otimes\mathbf{G})^\top(\mathbf{G}\otimes\mathbf{G})\mathbf{D}_m\beta = \mathbf{D}_m^\top(\mathbf{G}\otimes\mathbf{G})^\top\overline{\mathbf{y}}.$$

By using the identities

$$\mathbf{D}_m^\top vec(\mathbf{A}) = vech\left(\mathbf{A} + \mathbf{A}^\top - diag(\mathbf{A})\right)$$

and (A.54), said normal equation can be rewritten

$$vech\left(\mathbf{G}^\top\mathbf{G}\left(\mathbf{\Psi}+\mathbf{\Psi}^\top\right)\mathbf{G}^\top\mathbf{G} - diag\left(\mathbf{G}^\top\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\mathbf{G}\right)\right)$$
$$= vech\left(\mathbf{G}^\top\left(\overline{\mathbf{Y}}+\overline{\mathbf{Y}}^\top\right)\mathbf{G} - diag\left(\mathbf{G}^\top\overline{\mathbf{Y}}\mathbf{G}\right)\right).$$

Finally, it can be verified that $\widehat{\mathbf{\Psi}}$ given by (3.11) satisfies this equation as a consequence of the fact that such $\widehat{\mathbf{\Psi}}$ it holds that

$$vech\left(\mathbf{G}^\top\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\mathbf{G}\right) = vech\left(\mathbf{G}^\top\left(\frac{\overline{\mathbf{Y}}+\overline{\mathbf{Y}}^\top}{2}\right)\mathbf{G}\right).$$

$b)$ It straightforwardly follows from part $a)$. □

3.6.2 Proofs of main results

Proof of Proposition (3.2).

The proof follows the guidelines of the proof in Baraud [7]. More generally we will prove that for any $\eta > 0$ and any sequence of positive numbers L_m, if the penalty function $pen : \mathcal{M} \longrightarrow \mathbb{R}_+$ is chosen to satisfy:

$$pen\left(m\right) = \left(1+\eta+L_m\right)\frac{\delta_m^2}{N}D_m \text{ for all } m \in \mathcal{M}, \tag{3.27}$$

then for each $x > 0$ and $p \geq 2$,

$$\mathbb{P}\left(\mathcal{H}\left(\mathbf{f}\right) \geq \left(1+\frac{2}{\eta}\right)\frac{x}{N}\delta_m^2\right) \leq c\left(p,\eta\right)\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p \sum_{m\in\mathcal{M}} \frac{1}{\delta_m^p}\frac{D_m \vee 1}{\left(L_m D_m + x\right)^{p/2}}, \tag{3.28}$$

where we have set $\mathcal{H}\left(\mathbf{f}\right) = \left[\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - \left(2-\frac{4}{\eta}\right)\inf_{m\in\mathcal{M}}\left\{d_N^2\left(\mathbf{f},\mathcal{L}_m\right)+pen\left(m\right)\right\}\right]_+$.

To obtain (3.21), take $\eta = \frac{\theta}{2} = L_m$. As for each $m \in \mathcal{M}$,

$$\begin{aligned} d_N^2\left(\mathbf{f},\mathcal{L}_m\right)+pen\left(m\right) &\leq d_N^2\left(\mathbf{f},\mathcal{L}_m\right)+\left(1+\theta\right)\frac{\delta_m^2}{N}D_m \\ &\leq \left(1+\theta\right)\left(d_N^2\left(\mathbf{f},\mathcal{L}_m\right)+\frac{\delta_m^2}{N}D_m\right), \end{aligned}$$

we get that for all $q > 0$,

$$\mathcal{H}^q\left(\mathbf{f}\right) \geq \left[\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - \left(2+\frac{8}{\theta}\right)\left(1+\theta\right)M_N^*\right]_+^q = \left[\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - K\left(\theta\right)M_N^*\right]_+^q, \tag{3.29}$$

where $K\left(\theta\right) = \left(2+\frac{8}{\theta}\right)\left(1+\theta\right)$. Since

$$\mathbb{E}\left(\mathcal{H}^q\left(\mathbf{f}\right)\right) = \int_0^\infty q u^{q-1}\mathbb{P}\left(\mathcal{H}\left(\mathbf{f}\right) > u\right)du,$$

we derive from (3.29) and (3.28) that for all $p > 2\left(1+q\right)$,

$$\mathbb{E}\left[\left(\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - K\left(\theta\right)M_N^*\right)_+^q\right] \leq \mathbb{E}\left(\mathcal{H}^q\left(\mathbf{f}\right)\right)$$

$$\begin{aligned} &\leq c\left(p,\theta\right)\left(1+\frac{4}{\theta}\right)^q \frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{N^q}\sum_{m\in\mathcal{M}}\frac{\delta_m^{2q}}{\delta_m^p}\int_0^\infty q x^{q-1}\left[\frac{D_m\vee 1}{\left(\frac{\theta}{2}D_m+x\right)^{p/2}}\wedge 1\right]dx \\ &\leq c'\left(p,q,\theta\right)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{N^q}\delta_{\text{sup}}^{2q}\left[\sum_{m\in\mathcal{M}}\delta_m^{-p}D_m^{-\left(p/2-1-q\right)}\right] \end{aligned}$$

using that $\mathbb{P}\left(\mathcal{H}\left(\mathbf{f}\right) > u\right) \leq 1$.

Indeed, for $m \in \mathcal{M}$ such that $D_m \geq 1$, using that $q-1-p/2<0$, we get

$$\begin{aligned}
\frac{\delta_m^{2q}}{\delta_m^p} \int_0^\infty q x^{q-1} \left[\frac{D_m \vee 1}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \wedge 1 \right] dx &\leq \delta_{\sup}^{2q} \delta_m^{-p} \int_0^\infty q x^{q-1} \left[\frac{D_m}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \right] dx \\
&= \delta_{\sup}^{2q} \delta_m^{-p} \left(\int_0^{D_m} q x^{q-1} \left[\frac{D_m}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \right] dx + \int_{D_m}^\infty q x^{q-1} \left[\frac{D_m}{\left(\frac{\theta}{2} D_m + x\right)^{p/2}} \right] dx \right) \\
&\leq \delta_{\sup}^{2q} \delta_m^{-p} \left(\frac{D_m}{\left(\frac{\theta}{2} D_m\right)^{p/2}} \int_0^{D_m} q x^{q-1} dx + D_m \int_{D_m}^\infty q x^{q-1} \left[\frac{1}{x^{p/2}} \right] dx \right) \\
&= \delta_{\sup}^{2q} \delta_m^{-p} \left(2^{p/2} \theta^{-p/2} D_m^{1-p/2} \int_0^{D_m} q x^{q-1} dx + D_m \int_{D_m}^\infty q x^{q-1-p/2} dx \right) \\
&= \delta_{\sup}^{2q} \delta_m^{-p} \left(2^{p/2} \theta^{-p/2} D_m^{1-p/2} \left[D_m^q \right] + D_m \left[\frac{q}{p/2-q} D_m^{q-p/2} \right] \right) \\
&= \delta_{\sup}^{2q} \delta_m^{-p} \left(2^{p/2} \theta^{-p/2} D_m^{1-p/2+q} + D_m^{1-p/2+q} \left[\frac{q}{p/2-q} \right] \right) \\
&= \delta_{\sup}^{2q} \delta_m^{-p} \left(D_m^{-(p/2-1-q)} \left[2^{p/2} \theta^{-p/2} + \frac{q}{p/2-q} \right] \right). \qquad (3.30)
\end{aligned}$$

Inequality (3.30) enables to conclude that (3.21) holds assuming (3.28).

We now turn to the proof of (3.28). Recall that, we identify the function $g : \mathcal{X} \to \mathbb{R}^k$ with vectors $\left(g\left(x_1\right) \ldots g\left(x_N\right)\right)^\top \in \mathbb{R}^{Nk}$ and we denote by $\langle a, b \rangle_N = \frac{1}{N} \sum_{i=1}^N a_i^\top b_i$ the inner product of $\mathbb{R}^{Nk}$ associated to the norm $\|.\|_N$, where $a = \left(a_1 \ldots a_N\right)^\top$ and $b = \left(b_1 \ldots b_N\right)^\top$ with $a_i, b_i \in \mathbb{R}^k$ for all $i = 1, ..., N$. For each $m \in \mathcal{M}$ we denote by $\mathbf{P}_m$ the orthogonal projector onto the linear space $\left\{ \left(g\left(x_1\right) \ldots g\left(x_N\right)\right)^\top : g \in \mathcal{L}_m \right\} \subset \mathbb{R}^{Nk}$. This linear space is also denoted by $\mathcal{L}_m$. From now on, the subscript m denotes any minimizer of the function $m' \to \|\mathbf{f} - \mathbf{P}_{m'} \mathbf{f}\|_N^2 + pen\left(m'\right)$, $m' \in \mathcal{M}_N$. For any $\mathbf{g} \in \mathbb{R}^{Nk}$ we define the least squares loss function by

$$\gamma_N\left(\mathbf{g}\right) = \|\mathbf{y} - \mathbf{g}\|_N^2 .$$

Using the definition of γ_N we have that for all $\mathbf{g} \in \mathbb{R}^{Nk}$,

$$\gamma_N\left(\mathbf{g}\right) = \|\mathbf{f} + \boldsymbol{\varepsilon} - \mathbf{g}\|_N^2 .$$

Then we derive that

$$\|\mathbf{f} - \mathbf{g}\|_N^2 = \gamma_N\left(\mathbf{f}\right) + 2\left\langle \mathbf{f} - \mathbf{y}, \boldsymbol{\varepsilon} \right\rangle_N + \|\boldsymbol{\varepsilon}\|_N^2$$

and therefore

$$\left\|\mathbf{f} - \widetilde{\mathbf{f}}\right\|_N^2 - \|\mathbf{f} - \mathbf{P}_m \mathbf{f}\|_N^2 = \gamma_N\left(\widetilde{\mathbf{f}}\right) - \gamma_N\left(\mathbf{P}_m \mathbf{f}\right) + 2\left\langle \widetilde{\mathbf{f}} - \mathbf{P}_m \mathbf{f}, \boldsymbol{\varepsilon} \right\rangle_N . \qquad (3.31)$$

By the definition of $\widetilde{\mathbf{f}}$, we know that

$$\gamma_N\left(\widetilde{\mathbf{f}}\right) + pen\left(\widehat{m}\right) \leq \gamma_N\left(\mathbf{g}\right) + pen\left(m\right)$$

for all $m \in \mathcal{M}$ and for all $\mathbf{g} \in \mathcal{L}_m$. Then

$$\gamma_N\left(\widetilde{\mathbf{f}}\right) - \gamma_N\left(\mathbf{P}_m\mathbf{f}\right) \leq pen\left(m\right) - pen\left(\widehat{m}\right). \tag{3.32}$$

So we get from (3.31) and (3.32) that

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq \left\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\right\|_N^2 + pen\left(m\right) - pen\left(\widehat{m}\right) \\ + 2\left\langle \mathbf{f}-\mathbf{P}_m\mathbf{f},\boldsymbol{\varepsilon}\right\rangle_N + 2\left\langle \mathbf{P}_{\widehat{m}}\mathbf{f}-\mathbf{f},\boldsymbol{\varepsilon}\right\rangle_N + 2\left\langle \widetilde{\mathbf{f}}-\mathbf{P}_{\widehat{m}}\mathbf{f},\boldsymbol{\varepsilon}\right\rangle_N.\end{aligned} \tag{3.33}$$

In the following we set for each $m' \in \mathcal{M}$,

$$\begin{aligned}\mathcal{B}_{m'} &= \left\{\mathbf{g}\in\mathcal{L}_{m'} : \left\|\mathbf{g}\right\|_N \leq 1\right\}, \\ G_{m'} &= \sup_{t\in\mathcal{B}_{m'}} \left\langle \mathbf{g},\boldsymbol{\varepsilon}\right\rangle_N = \left\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\right\|_N, \\ \mathbf{u}_{m'} &= \begin{cases}\frac{\mathbf{P}_{m'}\mathbf{f}-\mathbf{f}}{\left\|\mathbf{P}_{m'}\mathbf{f}-\mathbf{f}\right\|_N} & \text{if } \left\|\mathbf{P}_{m'}\mathbf{f}-\mathbf{f}\right\|_N \neq 0 \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

Since $\widetilde{\mathbf{f}} = \mathbf{P}_{\widehat{m}}\mathbf{f} + \mathbf{P}_{\widehat{m}}\boldsymbol{\varepsilon}$, (3.33) gives

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq \left\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\right\|_N^2 + pen\left(m\right) - pen\left(\widehat{m}\right) \\ + 2\left\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\right\|_N \left|\left\langle \mathbf{u}_m,\boldsymbol{\varepsilon}\right\rangle_N\right| + 2\left\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\right\|_N \left|\left\langle \mathbf{u}_{\widehat{m}},\boldsymbol{\varepsilon}\right\rangle_N\right| + 2G_{\widehat{m}}^2.\end{aligned} \tag{3.34}$$

Using repeatedly the following elementary inequality that holds for all positive numbers α, x, z

$$2xz \leq \alpha x^2 + \frac{1}{\alpha}z^2 \tag{3.35}$$

we get for any $m' \in \mathcal{M}$

$$2\left\|\mathbf{f}-\mathbf{P}_{m'}\mathbf{f}\right\|_N \left|\left\langle \mathbf{u}_{m'},\boldsymbol{\varepsilon}\right\rangle_N\right| \leq \alpha\left\|\mathbf{f}-\mathbf{P}_{m'}\mathbf{f}\right\|_N^2 + \frac{1}{\alpha}\left|\left\langle \mathbf{u}_{m'},\boldsymbol{\varepsilon}\right\rangle_N\right|^2. \tag{3.36}$$

By Pythagoras Theorem we have

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 &= \left\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\right\|_N^2 + \left\|\mathbf{P}_{\widehat{m}}\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \\ &= \left\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\right\|_N^2 + G_{\widehat{m}}^2.\end{aligned} \tag{3.37}$$

We derive from (3.34) and (3.36) that for any $\alpha > 0$:

$$\begin{aligned}\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \leq\ & \left\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\right\|_N^2 + \alpha\left\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\right\|_N^2 + \frac{1}{\alpha}\left\langle \mathbf{u}_m,\boldsymbol{\varepsilon}\right\rangle_N^2 \\ & +\alpha\left\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\right\|_N^2 + \frac{1}{\alpha}\left\langle \mathbf{u}_{\widehat{m}},\boldsymbol{\varepsilon}\right\rangle_N^2 + 2G_{\widehat{m}}^2 + pen\left(m\right) - pen\left(\widehat{m}\right).\end{aligned}$$

Now taking into account that by equation (3.37), $\|\mathbf{f}-\mathbf{P}_{\widehat{m}}\mathbf{f}\|_N^2 = \left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - G_{\widehat{m}}^2$, the above inequality is equivalent to:

$$\begin{aligned}(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 \;\leq\;& (1+\alpha)\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + \frac{1}{\alpha}\langle \mathbf{u}_m,\boldsymbol{\varepsilon}\rangle_N^2 + \frac{1}{\alpha}\langle \mathbf{u}_{\widehat{m}},\boldsymbol{\varepsilon}\rangle_N^2 \\ &+ (2-\alpha)\,G_{\widehat{m}}^2 + pen\,(m) - pen\,(\widehat{m})\,. \end{aligned} \tag{3.38}$$

We choose $\alpha = \frac{2}{2+\eta} \in \,]0,1[$, but for sake of simplicity we keep using the notation α. Let $\widetilde{p}_1$ and $\widetilde{p}_2$ be two functions depending on η mapping $\mathcal{M}$ into $\mathbb{R}_+$. They will be specified later to satisfy

$$pen\,(m') \geq (2-\alpha)\,\widetilde{p}_1\,(m') + \frac{1}{\alpha}\widetilde{p}_2\,(m') \;\; \forall (m') \in \mathcal{M}. \tag{3.39}$$

Since $\frac{1}{\alpha}\widetilde{p}_2\,(m') \leq pen\,(m')$ and $1+\alpha \leq 2$, we get from (3.38) and (3.39) that

$$\begin{aligned}(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 &\leq (1+\alpha)\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + pen\,(m) + \frac{1}{\alpha}\widetilde{p}_2\,(m) + (2-\alpha)\left(G_{\widehat{m}}^2 - \widetilde{p}_1\,(\widehat{m})\right) \\ &\quad + \frac{1}{\alpha}\left(\langle \mathbf{u}_{\widehat{m}},\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_m,\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(m)\right) \\ &\leq 2\left(\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + pen\,(m)\right) + (2-\alpha)\left(G_{\widehat{m}}^2 - \widetilde{p}_1\,(\widehat{m})\right) \\ &\quad + \frac{1}{\alpha}\left(\langle \mathbf{u}_{\widehat{m}},\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_m,\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(m)\right). \end{aligned} \tag{3.40}$$

As $\frac{2}{1-\alpha} = 2 + \frac{4}{\eta}$ we obtain that

$$\begin{aligned}(1-\alpha)\,\mathcal{H}\,(\mathbf{f}) &= \left\{(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - (1-\alpha)\left(2+\frac{4}{\eta}\right)\inf_{m'\in\mathcal{M}}\left(\|\mathbf{f}-\mathbf{P}_{m'}\mathbf{f}\|_N^2 + pen\,(m')\right)\right\}_+ \\ &= \left\{(1-\alpha)\left\|\mathbf{f}-\widetilde{\mathbf{f}}\right\|_N^2 - 2\left(\|\mathbf{f}-\mathbf{P}_m\mathbf{f}\|_N^2 + 2pen\,(m)\right)\right\}_+ \\ &\leq \left\{(2-\alpha)\left(G_{\widehat{m}}^2 - \widetilde{p}_1\,(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_{\widehat{m}},\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(\widehat{m})\right) + \frac{1}{\alpha}\left(\langle \mathbf{u}_m,\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(m)\right)\right\}_+ \end{aligned}$$

using that m minimizes the function $\|\mathbf{f}-\mathbf{P}_{m'}\|^2 + pen\,(m')$ and (3.40).

For any $x > 0$,

$$\begin{aligned}\mathbb{P}\left((1-\alpha)\,\mathcal{H}\,(\mathbf{f}) \geq \frac{x\delta_m^2}{N}\right) &\leq \mathbb{P}\left(\exists m' \in \mathcal{M} : (2-\alpha)\left(G_{m'}^2 - \widetilde{p}_1\,(m')\right) \geq \frac{x\delta_{m'}^2}{3N}\right) \\ &\quad + \mathbb{P}\left(\exists m' \in \mathcal{M} : \frac{1}{\alpha}\left(\langle \mathbf{u}_{m'},\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(m')\right) \geq \frac{x\delta_{m'}^2}{3N}\right) \\ &\leq \sum_{m'\in\mathcal{M}} \mathbb{P}\left((2-\alpha)\left(\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\|_N^2 - \widetilde{p}_1\,(m')\right) \geq \frac{x\delta_{m'}^2}{3N}\right) \\ &\quad + \sum_{m'\in\mathcal{M}} \mathbb{P}\left(\frac{1}{\alpha}\left(\langle \mathbf{u}_{m'},\boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2\,(m')\right) \geq \frac{x\delta_{m'}^2}{3N}\right) \\ &:= \sum_{m'\in\mathcal{M}} P_{1,m'}\,(x) + \sum_{m'\in\mathcal{M}} P_{2,m'}\,(x)\,. \end{aligned} \tag{3.41}$$

We first bound $P_{2,m'}(x)$. Let t be some positive number,

$$\mathbb{P}\left(\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right| \geq t\right) \leq t^{-p}\mathbb{E}\left(\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right|^p\right). \tag{3.42}$$

Since $\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N = \frac{1}{N}\sum_{i=1}^{N} \langle \mathbf{u}_{im'}, \boldsymbol{\varepsilon}_i\rangle_{\ell_2}$ with $\boldsymbol{\varepsilon}_i \in \mathbb{R}^k$ i.i.d. and with zero mean, then by Rosenthal's inequality we know that for some constant $c(p)$ that depends on p only

$$\begin{aligned} c^{-1}(p)\, N^p \mathbb{E}\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right|^p &\leq \sum_{i=1}^{N} \mathbb{E}\left|\langle \mathbf{u}_{im'}, \boldsymbol{\varepsilon}_i\rangle_{\ell_2}\right|^p + \left(\sum_{i=1}^{N} \mathbb{E}\left(\langle \mathbf{u}_{im'}, \boldsymbol{\varepsilon}_i\rangle_{\ell_2}^2\right)\right)^{\frac{p}{2}} \\ &\leq \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{u}_{im'}\right\|_{\ell_2}^p \left\|\boldsymbol{\varepsilon}_i\right\|_{\ell_2}^p + \left(\sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{u}_{im'}\right\|_{\ell_2}^2 \left\|\boldsymbol{\varepsilon}_i\right\|_{\ell_2}^2\right)^{\frac{p}{2}} \\ &= \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p \sum_{i=1}^{N} \left\|\mathbf{u}_{im'}\right\|_{\ell_2}^p + \left(\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^2\right)^{\frac{p}{2}} \left(\sum_{i=1}^{N} \left\|\mathbf{u}_{im'}\right\|_{\ell_2}^2\right)^{\frac{p}{2}}. \end{aligned} \tag{3.43}$$

Since $p \geq 2$, $\left(\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^2\right)^{\frac{1}{2}} \leq \left(\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p\right)^{\frac{1}{p}}$ and

$$\left(\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^2\right)^{\frac{p}{2}} \leq \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p. \tag{3.44}$$

Using also that by definition $\left\|\mathbf{u}_{m'}\right\|_N^2 = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{u}_{im'}\right\|_{\ell_2}^2 = 1$, then $\frac{\left\|\mathbf{u}_{im'}\right\|_{\ell_2}^2}{N} \leq 1$ and therefore $\frac{\left\|\mathbf{u}_{im'}\right\|_{\ell_2}}{N^{\frac{1}{2}}} \leq 1$. Thus

$$\sum_{i=1}^{N}\left\|\mathbf{u}_{im'}\right\|_{\ell_2}^p = N^{\frac{p}{2}}\sum_{i=1}^{N}\left(\frac{\left\|\mathbf{u}_{im'}\right\|_{\ell_2}}{N^{\frac{1}{2}}}\right)^p \leq N^{\frac{p}{2}}\sum_{i=1}^{N}\left(\frac{\left\|\mathbf{u}_{im'}\right\|_{\ell_2}}{N^{\frac{1}{2}}}\right)^2 = N^{\frac{p}{2}}\left\|\mathbf{u}_{m'}\right\|_N^2 = N^{\frac{p}{2}}. \tag{3.45}$$

We deduce from (3.43), (3.44) and (3.45) that

$$c^{-1}(p)\, N^p \mathbb{E}\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right|^p \leq \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p N^{\frac{p}{2}} + \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p N^{\frac{p}{2}}.$$

Then for some constant $c'(p)$ that only depends on p,

$$\mathbb{E}\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right|^p \leq c'(p)\, \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p N^{-\frac{p}{2}}.$$

By this last inequality and (3.42) we get that

$$\mathbb{P}\left(\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right| \geq t\right) \leq c'(p)\, \mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p N^{-\frac{p}{2}} t^{-p}. \tag{3.46}$$

Let υ be some positive number depending on η only to be chosen later. We take t such that $Nt^2 = \min\left(\upsilon, \frac{\alpha}{3}\right)(L_{m'}D_{m'} + x)\,\delta_{m'}^2$ and set $N\widetilde{p}_2(m') = \upsilon L_{m'}D_{m'}\delta_{m'}^2$. We get

$$\begin{aligned}
P_{2,m'}(x) &= \mathbb{P}\left(\frac{1}{\alpha}\left(\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N^2 - \widetilde{p}_2(m')\right) \geq \frac{x\delta_{m'}^2}{3N}\right) \\
&= \mathbb{P}\left(N\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N^2 \geq N\widetilde{p}_2(m') + \alpha\frac{\delta_{m'}^2}{3}x\right) \\
&= \mathbb{P}\left(N\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N^2 \geq \upsilon L_{m'}D_{m'}\delta_{m'}^2 + \alpha\frac{\delta_{m'}^2}{3}x\right) \\
&\leq \mathbb{P}\left(\left|\langle \mathbf{u}_{m'}, \boldsymbol{\varepsilon}\rangle_N\right| \geq N^{-\frac{1}{2}}\sqrt{\min\left(\upsilon, \frac{\alpha}{3}\right)}\sqrt{(L_{m'}D_{m'} + x)}\delta_{m'}\right) \\
&\leq c'(p)\,\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p N^{-\frac{p}{2}}\frac{N^{\frac{p}{2}}}{\left(\min\left(\upsilon, \frac{\alpha}{3}\right)\right)^{\frac{p}{2}}(L_{m'}D_{m'} + x)^{\frac{p}{2}}\delta_{m'}^p} \\
&= c''(p,\eta)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}\frac{1}{(L_{m'}D_{m'} + x)^{\frac{p}{2}}}. \qquad (3.47)
\end{aligned}$$

The last inequality holds using (3.46).

We now bound $P_{1,m'}(x)$ for those $m' \in \mathcal{M}$ such that $D_{m'} \geq 1$. By using our version of Corollary 5.1 in Baraud with $\widetilde{\mathbf{A}} = \mathbf{P}_{m'}$, $\operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) = D_{m'}$ and $\tau\left(\widetilde{\mathbf{A}}\right) = 1$, we obtain from (3.22) that for any positive $x_{m'}$

$$\mathbb{P}\left(N\left\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\right\|_N^2 \geq \delta_{m'}^2 D_{m'} + 2\delta_{m'}^2\sqrt{D_{m'}x_{m'}} + \delta_{m'}^2 x_{m'}\right) \leq C(p)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}D_{m'}x_{m'}^{-\frac{p}{2}}. \qquad (3.48)$$

Since for any $\beta > 0$, $2\sqrt{D_{m'}x_{m'}} \leq \beta D_{m'} + \beta^{-1}x_{m'}$ then (3.48) imply that

$$\mathbb{P}\left(N\left\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\right\|_N^2 \geq (1+\beta)D_{m'}\delta_{m'}^2 + \left(1+\beta^{-1}\right)x_{m'}\delta_{m'}^2\right) \leq C(p)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}D_{m'}x_{m'}^{-\frac{p}{2}}. \qquad (3.49)$$

Now for some number β depending on η only to be chosen later, we take

$$x_{m'} = \left(1+\beta^{-1}\right)\min\left(\upsilon, \frac{(2-\alpha)^{-1}}{3}\right)(L_{m'}D_{m'} + x)$$

and $N\widetilde{p}_1(m') = \upsilon L_{m'}D_{m'}\delta_{m'}^2 + (1+\beta)D_{m'}\delta_{m'}^2$. By (3.49) this gives

$$\begin{aligned}
P_{1,m'}(x) &= \mathbb{P}\left(\left\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\right\|_N^2 - \widetilde{p}_1(m') \geq \frac{(2-\alpha)^{-1}x\delta_{m'}^2}{3N}\right) \\
&= \mathbb{P}\left(N\left\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\right\|_N^2 \geq \upsilon L_{m'}D_{m'}\delta_{m'}^2 + (1+\beta)D_{m'}\delta_{m'}^2 + \frac{(2-\alpha)^{-1}}{3}x\delta_{m'}^2\right) \\
&\leq \mathbb{P}\left(N\left\|\mathbf{P}_{m'}\boldsymbol{\varepsilon}\right\|_N^2 \geq (1+\beta)D_{m'}\delta_{m'}^2 + \left(1+\beta^{-1}\right)x_{m'}\delta_{m'}^2\right) \\
&\leq c(p)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}D_{m'}x_{m'}^{-\frac{p}{2}} \\
&\leq c'(p,\eta)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}\frac{D_{m'}}{(L_{m'}D_{m'} + x)^{\frac{p}{2}}}. \qquad (3.50)
\end{aligned}$$

Gathering (3.47), (3.50) and (3.41) we get that

$$\begin{aligned}\mathbb{P}\left(\mathcal{H}(\mathbf{f}) \geq \frac{x\delta_{m'}^2}{N(1-\alpha)}\right) &\leq \sum_{m'\in\mathcal{M}} P_{1,m'}(x) + \sum_{m'\in\mathcal{M}} P_{2,m'}(x)\\ &\leq \sum_{m'\in\mathcal{M}} c'(p,\eta)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}\frac{D_{m'}}{(L_{m'}D_{m'}+x)^{\frac{p}{2}}}\\ &+\sum_{m'\in\mathcal{M}} c''(p,\eta)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}\frac{1}{(L_{m'}D_{m'}+x)^{\frac{p}{2}}}.\end{aligned}$$

Since $\frac{1}{(1-\alpha)} = (1+2\eta^{-1})$, then (3.28) holds:

$$\begin{aligned}\mathbb{P}\left(\mathcal{H}(\mathbf{f}) \geq (1+2\eta^{-1})\frac{x\delta_{m'}^2}{N}\right) &\leq \sum_{m'\in\mathcal{M}} \frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p (L_{m'}D_{m'}+x)^{\frac{p}{2}}}\max(D_{m'},1)\left(c'(p,\eta)+c''(p,\eta)\right)\\ &= c(p,\eta)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta_{m'}^p}\sum_{m'\in\mathcal{M}}\frac{D_{m'}\vee 1}{(L_{m'}D_{m'}+x)^{\frac{p}{2}}}.\end{aligned}$$

It remains to choose β and δ for (3.39) to hold (we recall that $\alpha = \frac{2}{2+\eta}$). This is the case if $(2-\alpha)(1+\beta) = 1+\eta$ and $(2-\alpha+\alpha^{-1})\delta = 1$, therefore we take $\beta = \frac{\eta}{2}$ and $\delta = \left[1+\frac{\eta}{2}+2\frac{(1+\eta)}{(2+\eta)}\right]^{-1}$. □

3.6.3 Proof of the concentration inequality

Proof of Proposition (3.3).

Since $\widetilde{\mathbf{A}}$ is non-negative definite and symmetric there exists $\mathbf{A} \in \mathbb{R}^{Nk\times Nk}\backslash\{\mathbf{0}\}$ such that $\widetilde{\mathbf{A}} = \mathbf{A}^\top\mathbf{A}$. Then

$$\begin{aligned}\zeta^2(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^\top\widetilde{\mathbf{A}}\boldsymbol{\varepsilon} = (\mathbf{A}\boldsymbol{\varepsilon})^\top\mathbf{A}\boldsymbol{\varepsilon} = \left\|\mathbf{A}\boldsymbol{\varepsilon}\right\|_{\ell_2}^2 &= \left[\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1}\left\langle\mathbf{A}\boldsymbol{\varepsilon},\mathbf{u}\right\rangle_{\ell_2}\right]^2\\ = \left[\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1}\left\langle\boldsymbol{\varepsilon},\mathbf{A}^\top\mathbf{u}\right\rangle_{\ell_2}\right]^2 &= \left[\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1}\sum_{i=1}^N\left\langle\boldsymbol{\varepsilon}_i,\left(\mathbf{A}^\top\mathbf{u}\right)_i\right\rangle_{\ell_2}\right]^2\\ = \left[\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1}\sum_{i=1}^N\left\langle\boldsymbol{\varepsilon}_i,\mathbf{A}_i^\top\mathbf{u}\right\rangle_{\ell_2}\right]^2 &= \left[\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1}\sum_{i=1}^N\sum_{j=1}^k\boldsymbol{\varepsilon}_{ij}\left(\mathbf{A}_i^\top\mathbf{u}\right)_j\right]^2\end{aligned}$$

with $\mathbf{A} = (\mathbf{A}_1 \mid ... \mid \mathbf{A}_N)$, where $\mathbf{A}_i$ is an $Nk\times k$ matrix. Now take

$$\mathcal{G} = \left\{\begin{array}{c} g_{\mathbf{u}} : g_{\mathbf{u}}(\mathbf{x}) = \sum\limits_{i=1}^N\left\langle\mathbf{x}_i,\mathbf{A}_i^\top\mathbf{u}\right\rangle_{\ell_2} = \sum\limits_{i=1}^N\left\langle\mathbf{B}_i\mathbf{x},\mathbf{B}_i\mathbf{A}^\top\mathbf{u}\right\rangle_{\ell_2},\\ \mathbf{u},\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^\top\in\mathbb{R}^{Nk},\ \left\|\mathbf{u}\right\|_{\ell_2}\leq 1\end{array}\right\},$$

where $\mathbf{B}_i = [\mathbf{0},...,\mathbf{0},\mathbf{I}_k,\mathbf{0},...\mathbf{0}] \in \mathbb{R}^{k\times Nk}$ for $i = 1,...,N$, and $\mathbf{I}_k$ is the identity matrix in $\mathbb{R}^{k\times k}$. Let $\mathbf{M}_i = [\mathbf{0},\ldots,\mathbf{0},\mathbf{I}_k,\mathbf{0},\ldots,\mathbf{0}]^\top \in \mathbb{R}^{Nk\times Nk}$ for $i = 1,...,N$, such that $\boldsymbol{\varepsilon}_i = \mathbf{B}_i\boldsymbol{\varepsilon}$

and $\mathbf{M}_i\boldsymbol{\varepsilon} = [\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\varepsilon}_i, \mathbf{0}, \ldots, \mathbf{0}]^\top$. Then

$$\zeta(\boldsymbol{\varepsilon}) = \sup_{\|\mathbf{u}\|_{\ell_2} \leq 1} \sum_{i=1}^{N} g_{\mathbf{u}}(\mathbf{M}_i\boldsymbol{\varepsilon}).$$

Now take $\mathbf{U}_i = \mathbf{M}_i\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^{Nk}$. Then for each positive number t and $p > 0$,

$$\begin{aligned}
\mathbb{P}(\zeta(\boldsymbol{\varepsilon}) \geq \mathbb{E}(\zeta(\boldsymbol{\varepsilon})) + t) &\leq \mathbb{P}(|\zeta(\boldsymbol{\varepsilon}) - \mathbb{E}(\zeta(\boldsymbol{\varepsilon}))| > t) \\
&\leq t^{-p}\mathbb{E}(|\zeta(\boldsymbol{\varepsilon}) - \mathbb{E}(\zeta(\boldsymbol{\varepsilon}))|^p) \text{ by Markov inequality} \\
&\leq c(p)\,t^{-p} \left\{ \begin{array}{c} \mathbb{E}\left(\max\limits_{i=1,\ldots,N} \sup\limits_{\|\mathbf{u}\|_{\ell_2} \leq 1} \left| \left\langle \boldsymbol{\varepsilon}_i, \mathbf{A}_i^\top \mathbf{u} \right\rangle_{\ell_2} \right|^p \right) \\ + \left[\mathbb{E}\left(\sup\limits_{\|\mathbf{u}\|_{\ell_2} \leq 1} \sum\limits_{i=1}^{N} \left(\left\langle \boldsymbol{\varepsilon}_i, \mathbf{A}_i^\top \mathbf{u} \right\rangle_{\ell_2} \right)^2 \right) \right]^{p/2} \end{array} \right\} \\
&= c(p)\,t^{-p} \left(\mathbb{E}_1 + \mathbb{E}_2^{p/2} \right). \qquad (3.51)
\end{aligned}$$

We start by bounding $\mathbb{E}_1$. For all $\mathbf{u}$ such that $\|\mathbf{u}\|_{\ell_2} \leq 1$ and $i \in \{1, ..., N\}$,

$$\left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 \leq \left\|\mathbf{A}^\top \mathbf{u}\right\|_{\ell_2}^2 \leq \|\mathbf{A}\|_2^2,$$

where $\|\mathbf{A}\|_2 = \sup\limits_{\mathbf{x} \in \mathbb{R}^{Nk}:\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_{\ell_2}}{\|\mathbf{x}\|_{\ell_2}}$ for all matrix $\mathbf{A}$. For $p \geq 2$ the following inequality holds

$$\left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^p \leq \|\mathbf{A}\|_2^{p-2} \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2,$$

then

$$\left| \left\langle \boldsymbol{\varepsilon}_i, \mathbf{A}_i^\top \mathbf{u} \right\rangle_{\ell_2} \right|^p \leq \left[\|\boldsymbol{\varepsilon}_i\|_{\ell_2} \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2} \right]^p \leq \|\mathbf{A}\|_2^{p-2} \|\boldsymbol{\varepsilon}_i\|_{\ell_2}^p \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2.$$

Therefore

$$\mathbb{E}_1 \leq \|\mathbf{A}\|_2^{p-2}\, \mathbb{E}\left(\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{N} \|\boldsymbol{\varepsilon}_i\|_{\ell_2}^p \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 \right).$$

Since $\|\mathbf{u}\|_{\ell_2} \leq 1$, $\forall i = 1, ..., N$

$$\left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 = \mathbf{u}^\top \mathbf{A}_i \mathbf{A}_i^\top \mathbf{u} \leq \left\|\mathbf{A}_i \mathbf{A}_i^\top\right\|_2 \leq \operatorname{Tr}\left(\mathbf{A}_i \mathbf{A}_i^\top\right),$$

then

$$\sum_{i=1}^{N} \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 \leq \sum_{i=1}^{N} \operatorname{Tr}\left(\mathbf{A}_i \mathbf{A}_i^\top\right) = \operatorname{Tr}\left(\sum_{i=1}^{N} \mathbf{A}_i \mathbf{A}_i^\top \right) = \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right).$$

Thus,

$$\mathbb{E}_1 \leq \|\mathbf{A}\|_2^{p-2} \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right). \qquad (3.52)$$

We now bound $\mathbb{E}_2$ via a truncation argument. Since for all $\mathbf{u}$ such that $\|\mathbf{u}\|_{\ell_2} \leq 1$, $\left\|\mathbf{A}^\top \mathbf{u}\right\|_{\ell_2}^2 \leq \|\mathbf{A}\|_2^2$, for any positive number c to be specified later we have that

$$\begin{aligned}
\mathbb{E}_2 &\leq \mathbb{E}\left(\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1} \sum_{i=1}^N \|\boldsymbol{\varepsilon}_i\|_{\ell_2}^2 \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 1_{\left\{\|\boldsymbol{\varepsilon}_i\|_{\ell_2}\leq c\right\}} \right) + \mathbb{E}\left(\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1} \sum_{i=1}^N \|\boldsymbol{\varepsilon}_i\|_{\ell_2}^2 \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 1_{\left\{\|\boldsymbol{\varepsilon}_i\|_{\ell_2}> c\right\}} \right) \\
&\leq \mathbb{E}\left(c^2 \sup_{\|\mathbf{u}\|_{\ell_2}\leq 1} \sum_{i=1}^N \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 1_{\left\{\|\boldsymbol{\varepsilon}_i\|_{\ell_2}\leq c\right\}} \right) + \mathbb{E}\left(\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1} \sum_{i=1}^N \|\boldsymbol{\varepsilon}_i\|_{\ell_2}^2 \left\|\mathbf{A}_i^\top \mathbf{u}\right\|_{\ell_2}^2 1_{\left\{\|\boldsymbol{\varepsilon}_i\|_{\ell_2}> c\right\}} \right) \\
&\leq c^2 \|\mathbf{A}\|_2^2 + c^{2-p}\mathbb{E}\left(\sup_{\|\mathbf{u}\|_{\ell_2}\leq 1} \sum_{i=1}^N \|\mathbf{A}_i \mathbf{u}\|_{\ell_2}^2 \|\boldsymbol{\varepsilon}_i\|_{\ell_2}^p \right) \\
&\leq c^2 \|\mathbf{A}\|_2^2 + c^{2-p}\mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \qquad (3.53)
\end{aligned}$$

using the bound obtained for $\mathbb{E}_1$. It remains to take $c^p = \mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) / \|\mathbf{A}\|_2^2$ to get that:

$$\mathbb{E}_2 \leq c^2 \|\mathbf{A}\|_2^2 + c^2 \|\mathbf{A}\|_2^2 = 2c^2 \|\mathbf{A}\|_2^2 ,$$

therefore

$$\mathbb{E}_2^{p/2} \leq 2^{p/2} c^p \|\mathbf{A}\|_2^p , \qquad (3.54)$$

which implies that

$$2^{-p/2}\mathbb{E}_2^{p/2} \leq \mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \|\mathbf{A}\|_2^{p-2} .$$

We straightforwardly derive from (3.51) that

$$\mathbb{P}\left(\zeta^2(\boldsymbol{\varepsilon}) \geq \left[\mathbb{E}(\zeta(\boldsymbol{\varepsilon}))\right]^2 + 2\mathbb{E}(\zeta(\boldsymbol{\varepsilon}))\, t + t^2\right) \leq c(p)\, t^{-p} \left(\mathbb{E}_1 + \mathbb{E}_2^{p/2}\right) .$$

Since $\left[\mathbb{E}(\zeta(\boldsymbol{\varepsilon}))\right]^2 \leq \mathbb{E}\left(\zeta^2(\boldsymbol{\varepsilon})\right)$, (3.52) and (3.54) imply that

$$\begin{aligned}
&\mathbb{P}\left(\zeta^2(\boldsymbol{\varepsilon}) \geq \mathbb{E}\left(\zeta^2(\boldsymbol{\varepsilon})\right) + 2\sqrt{\mathbb{E}\left(\zeta^2(\boldsymbol{\varepsilon})\right) t^2} + t^2\right) \leq c(p)\, t^{-p} \left(\mathbb{E}_1 + \mathbb{E}_2^{p/2}\right) \\
&\leq c(p)\, t^{-p} \left(\|\mathbf{A}\|_2^{p-2} \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) + 2^{p/2}\mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \|\mathbf{A}\|_2^{p-2} \right) \\
&\leq c'(p)\, t^{-p} \|\mathbf{A}\|_2^{p-2} \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) , \qquad (3.55)
\end{aligned}$$

for all $t > 0$. Moreover

$$\begin{aligned}
\mathbb{E}\left(\zeta^2(\boldsymbol{\varepsilon})\right) &= \mathbb{E}\left(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{A}} \boldsymbol{\varepsilon}\right) = \mathbb{E}\left(\operatorname{Tr}\left(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{A}} \boldsymbol{\varepsilon}\right)\right) = \mathbb{E}\left(\operatorname{Tr}\left(\widetilde{\mathbf{A}} \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right)\right) \\
&= \operatorname{Tr}\left(\widetilde{\mathbf{A}}\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right)\right) = \operatorname{Tr}\left(\widetilde{\mathbf{A}}\left(\mathbf{I}_N \otimes \boldsymbol{\Phi}\right)\right) = \delta^2 \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) . \qquad (3.56)
\end{aligned}$$

Using (3.56), take $t^2 = \tau\left(\widetilde{\mathbf{A}}\right) \delta^2 x > 0$ in (3.55) to get that

$$\begin{aligned}
&\mathbb{P}\left(\zeta^2(\boldsymbol{\varepsilon}) \geq \delta^2 \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) + 2\sqrt{\delta^2 \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \tau\left(\widetilde{\mathbf{A}}\right) \delta^2 x} + \tau\left(\widetilde{\mathbf{A}}\right) \delta^2 x \right) \\
\leq\ & c'(p)\, \tau^{-p/2}\left(\widetilde{\mathbf{A}}\right) \delta^{-p/2} x^{-p/2} \|\mathbf{A}\|_2^{p-2} \operatorname{Tr}\left(\widetilde{\mathbf{A}}\right) \mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) ,
\end{aligned}$$

where $\tau\left(\widetilde{\mathbf{A}}\right)$ is the spectral radius of $\widetilde{\mathbf{A}}$. Since $\widetilde{\mathbf{A}}$ is a non-negative definite matrix then $\tau\left(\widetilde{\mathbf{A}}\right) = \tau\left(\mathbf{A}^\top\mathbf{A}\right) = \left\|\mathbf{A}^\top\mathbf{A}\right\|_2 = \|\mathbf{A}\|_2^2$ (see definitions A.7, A.8 and property A.3 in the Appendix). Hence, the desired result follows:

$$\begin{aligned} & \mathbb{P}\left(\zeta^2\left(\boldsymbol{\varepsilon}\right) \geq \delta^2 \mathrm{Tr}\left(\widetilde{\mathbf{A}}\right) + 2\delta^2\sqrt{\tau\left(\widetilde{\mathbf{A}}\right)\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)x} + \delta^2\tau\left(\widetilde{\mathbf{A}}\right)x\right) \\ \leq \; & c'\left(p\right)\tau^{-p/2}\left(\widetilde{\mathbf{A}}\right)\delta^{-p/2}x^{-p/2}\tau^{(p-2)/2}\left(\widetilde{\mathbf{A}}\right)\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)\mathbb{E}\left(\|\boldsymbol{\varepsilon}_1\|_{\ell_2}^p\right) \\ = \; & c'\left(p\right)\frac{\mathbb{E}\left\|\boldsymbol{\varepsilon}_1\right\|_{\ell_2}^p}{\delta^p}\frac{\mathrm{Tr}\left(\widetilde{\mathbf{A}}\right)}{\tau\left(\widetilde{\mathbf{A}}\right)x^{p/2}}. \end{aligned}$$

□

Chapter 4

Group-Lasso estimation of high-dimensional covariance matrices

Abstract: In this chapter, we consider the Group-Lasso estimator of the covariance matrix of a stochastic process corrupted by an additive noise. We propose to estimate the covariance matrix in a high-dimensional setting under the assumption that the process has a sparse representation in a large dictionary of basis functions. Using a matrix regression model, we propose a new methodology for high-dimensional covariance matrix estimation based on empirical contrast regularization by a Group-Lasso penalty. Using such a penalty, the method selects a sparse set of basis functions in the dictionary used to approximate the process, leading to an approximation of the covariance matrix into a low dimensional space. Consistency of the estimator is studied in Frobenius and operator norms and an application to sparse PCA is proposed.

4.1 Introduction

Let T be some subset of $\mathbb{R}^d$, $d \in \mathbb{N}$, and let $X = \{X(t) : t \in T\}$ be a stochastic process with values in $\mathbb{R}$. Assume that X has zero mean $\mathbb{E}(X(t)) = 0$ for all $t \in T$, and finite covariance $\sigma(s,t) = \mathbb{E}(X(s)X(t))$ for all $s, t \in T$. Let $t_1, \ldots, t_n$ be fixed points in T (deterministic design), $X_1, ..., X_N$ independent copies of the process X, and suppose that we observe the noisy processes

$$\widetilde{X}_i(t_j) = X_i(t_j) + \mathcal{E}_i(t_j) \text{ for } i = 1, ..., N,\ j = 1, ..., n, \tag{4.1}$$

where $\mathcal{E}_1, ..., \mathcal{E}_N$ are independent copies of a second order Gaussian process $\mathcal{E}$ with zero mean and independent of X, which represent an additive source of noise in the measurements. Based on the noisy observations (4.1), an important problem in statistics is to construct an estimator of the covariance matrix $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ of the process X at the design points, where $\mathbf{X} = (X(t_1), ..., X(t_n))^\top$. This problem is a fundamental issue in many statistical applications. For instance, estimating such a covariance matrix has important applications in dimension reduction by principal component analysis (PCA) or classification by linear or quadratic discriminant analysis (LDA and QDA).

In Bigot, Biscay, Loubes and Muñiz [14], using N independent copies of the process X, we have proposed to construct an estimator of the covariance matrix $\mathbf{\Sigma}$ by expanding the process X into a dictionary of basis functions. The method in Bigot et al. [14] is based on model selection techniques by empirical contrast minimization in a suitable matrix regression model. This new approach to covariance estimation is well adapted to the case of low-dimensional covariance estimation when the number of replicates N of the process is larger than the number of observations points n. However, many application areas are currently dealing with the problem of estimating a covariance matrix when the number of observations at hand is small when compared to the number of parameters to estimate. Examples include biomedical imaging, proteomic/genomic data, signal processing in neurosciences and many others. This issue corresponds to the problem of covariance estimation for high-dimensional data. This problem is challenging since, in a high-dimensional setting (when $n >> N$ or $n \sim N$), it is well known that the sample covariance matrices

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^\top \in \mathbb{R}^{n\times n},\ \text{where}\ \ \mathbf{X}_i = \left(X_i\left(t_1\right), ..., X_i\left(t_n\right)\right)^\top, i = 1, \ldots, N$$

and

$$\widetilde{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top \in \mathbb{R}^{n\times n},\ \text{where}\ \ \widetilde{\mathbf{X}}_i = \left(\widetilde{X}_i\left(t_1\right), ..., \widetilde{X}_i\left(t_n\right)\right)^\top, i = 1, \ldots, N$$

behave poorly, and are not consistent estimators of $\mathbf{\Sigma}$. For example, suppose that the $\mathbf{X}_i$'s are independent and identically distributed (i.i.d.) random vectors in $\mathbb{R}^n$ drawn from a multivariate Gaussian distribution. Then, when $\frac{n}{N} \to c > 0$ as $n, N \to +\infty$, neither the eigenvalues nor the eigenvectors of the sample covariance matrix $\mathbf{S}$ are consistent estimators of the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ (see Johnstone [49]). This topic has thus recently received a lot of attention in the statistical literature. To achieve consistency, recently developed methods for high-dimensional covariance estimation impose sparsity restrictions on the matrix $\mathbf{\Sigma}$. Such restrictions imply that the true (but unknown) dimension of the model is much lower than the number $\frac{n(n+1)}{2}$ of parameters of an unconstrained covariance matrix. Under various sparsity assumptions, different regularizing methods of the empirical covariance matrix have been proposed. Estimators based on thresholding or banding the entries of the empirical covariance matrix have been studied in Bickel and Levina [11] and Bickel and Levina [12]. Thresholding the components of the empirical covariance matrix has also been proposed by El Karoui [38] and the consistency of such estimators is studied using tools from random matrix theory. Fan, Fan and Lv [41] impose sparsity on the covariance via a factor model which is appropriate in financial applications. Levina, Rothman and Zhu [57] and Rothman, Bickel, Levina and Zhu [79] propose regularization techniques with a Lasso penalty to estimate the covariance matrix or its inverse. More general penalties have been studied in Lam and Fan [56]. Another approach is to impose sparsity on the eigenvectors of the covariance matrix which leads to sparse PCA. Zou, Hastie and Tibshirani [93] use a Lasso penalty to achieve sparse representation in PCA, d'Aspremont, Bach and El Ghaoui [30] study properties of sparse principal components by convex programming, while Johnstone and Lu [50] propose a

PCA regularization by expanding the empirical eigenvectors in a sparse basis and then apply a thresholding step.

In this chapter, we propose to estimate $\mathbf{\Sigma}$ in a high-dimensional setting by using the assumption that the process X has a sparse representation in a large dictionary of basis functions. Using a matrix regression model as in Bigot et al. [14], we propose a new methodology for high-dimensional covariance matrix estimation based on empirical contrast regularization by a Group-Lasso penalty. Using such a penalty, the method selects a sparse set of basis functions in the dictionary used to approximate the process X. This leads to an approximation of the covariance matrix $\mathbf{\Sigma}$ into a low dimensional space, and thus to a new method of dimension reduction for high-dimensional data. Group-Lasso estimators have been studied in the standard linear model and in multiple kernel learning to impose a group-sparsity structure on the parameters to recover (see Nardi and Rinaldo [70], Bach [5] and references therein). However, to the best of our knowledge, it has not been used for the estimation of covariance matrices using a functional approximation of the process X.

The rest of the chapter is organized as follows. In Section 4.2, we describe a matrix regression model for covariance estimation, and we define our estimator by Group-Lasso regularization. The consistency of such a procedure is investigated in Section 4.3 using oracle inequalities and a non-asymptotic point of view by holding fixed the number of replicates N and observation points n. Consistency of the estimator is studied in Frobenius and operator norms. Various results existing in matrix theory show that convergence in operator norm implies convergence of the eigenvectors and eigenvalues (see e.g. El Karoui [38] and references therein). Consistency in operator norm is thus well suited for PCA applications. Numerical experiments are given in Section 4.4, and an application to sparse PCA is proposed. All the proofs are contained in Section 4.5 at the end of the chapter.

4.2 Model and definition of the estimator

To impose sparsity restrictions on the covariance matrix $\mathbf{\Sigma}$, our approach is based on an approximation of the process in a finite dictionary of (not necessarily orthogonal) basis functions $g_m : T \to \mathbb{R}$ for $m = 1, ..., M$. Suppose that

$$X(t) \approx \sum_{m=1}^{M} a_m g_m(t), \tag{4.2}$$

where a_m, $m = 1, ..., M$ are real valued random variables, and that for each trajectory X_i,

$$X_i(t_j) \approx \sum_{m=1}^{M} a_{i,m} g_m(t_j). \tag{4.3}$$

The notation $\approx$ means that the process X can be well approximated into the dictionary. A precise meaning of this will be discussed later on. Then (4.3) can be written in matrix notation as:

$$\mathbf{X}_i \approx \mathbf{G}\mathbf{a}_i, \; i = 1, ..., N, \tag{4.4}$$

where $\mathbf{G}$ is the $n \times M$ matrix with entries

$$\mathbf{G}_{jm} = g_m(t_j) \text{ for } 1 \leq j \leq n \text{ and } 1 \leq m \leq M,$$

and $\mathbf{a}_i$ is the $M \times 1$ random vector of components $a_{i,m}$, with $1 \leq m \leq M$.

Recall that we want to estimate the covariance matrix $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ from the noisy observations (4.1). Since $\mathbf{X} \approx \mathbf{G}\mathbf{a}$, where $\mathbf{a} = (a_m)_{1 \leq m \leq M}$ with a_m as in (4.2), it follows that

$$\mathbf{\Sigma} \approx \mathbb{E}\left(\mathbf{G}\mathbf{a}\left(\mathbf{G}\mathbf{a}\right)^\top\right) = \mathbb{E}\left(\mathbf{G}\mathbf{a}\mathbf{a}^\top\mathbf{G}^\top\right) = \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top \text{ with } \mathbf{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right).$$

Given the noisy observations $\widetilde{\mathbf{X}}_i$ as in (4.1) with $i = 1, ..., N$, consider the following matrix regression model

$$\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top = \mathbf{\Sigma} + \mathbf{U}_i + \mathbf{W}_i \; i = 1, \ldots, N,$$

where $\mathbf{U}_i = \mathbf{X}_i\mathbf{X}_i^\top - \mathbf{\Sigma}$ are i.i.d. centered matrix errors, and

$$\mathbf{W}_i = \mathcal{E}_i\mathcal{E}_i^\top \in \mathbb{R}^{n\times n} \text{ where } \mathcal{E}_i = \left(\mathcal{E}_i(t_1), ..., \mathcal{E}_i(t_n)\right)^\top, i = 1, \ldots, N.$$

The size M of the dictionary can be very large, but it is expected that the process X has a sparse expansion in this basis, meaning that, in approximation (4.2), many of the random coefficients a_m are close to zero. We are interested in obtaining an estimate of the covariance $\mathbf{\Sigma}$ in the form $\widehat{\mathbf{\Sigma}} = \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top$ such that $\widehat{\mathbf{\Psi}}$ is a symmetric $M \times M$ matrix with many zero rows (and so, by symmetry, many corresponding zero columns). Note that setting the k-th row of $\widehat{\mathbf{\Psi}}$ to $\mathbf{0} \in \mathbb{R}^M$ means to remove the function g_k from the set of basis functions $(g_m)_{1 \leq m \leq M}$ in the function expansion associated to $\mathbf{G}$.

Let us now explain how to select a sparse set of rows/columns in the matrix $\widehat{\mathbf{\Psi}}$. For this, we use a Group-Lasso approach to threshold some rows/columns of $\widehat{\mathbf{\Psi}}$ which corresponds to removing some basis functions in the approximation of the process X. Let $\mathcal{S}_M$ denote the set of $M \times M$ symmetric matrices with real entries. We define the Group-Lasso estimator of the covariance matrix $\mathbf{\Sigma}$ by

$$\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}_\lambda\mathbf{G}^\top \in \mathbb{R}^{n\times n}, \tag{4.5}$$

where $\widehat{\mathbf{\Psi}}_\lambda$ is the solution of the following optimization problem:

$$\widehat{\mathbf{\Psi}}_\lambda = \underset{\mathbf{\Psi}\in\mathcal{S}_M}{\operatorname{argmin}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \left\| \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|_F^2 + 2\lambda \sum_{k=1}^{M} \gamma_k \sqrt{\sum_{m=1}^{M} \Psi_{mk}^2} \right\}, \tag{4.6}$$

where $\mathbf{\Psi} = (\Psi_{mk})_{1\leq m,k \leq M} \in \mathbb{R}^{M\times M}$, λ is a positive number and γ_k are some weights whose values will be discuss later on. In (4.6), the penalty term imposes to give preference to solutions with components $\mathbf{\Psi}_k = \mathbf{0}$, where $(\mathbf{\Psi}_k)_{1\leq k\leq M}$ denotes the columns of $\mathbf{\Psi}$. Recall that $\widetilde{\mathbf{S}} = \frac{1}{N}\sum\limits_{i=1}^{N} \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top$ denotes the sample covariance matrix from the noisy observations (4.1). It can be checked that minimizing the criterion (4.6) is equivalent to

$$\widehat{\mathbf{\Psi}}_\lambda = \underset{\mathbf{\Psi}\in\mathcal{S}_M}{\operatorname{argmin}} \left\{ \left\| \widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top \right\|_F^2 + 2\lambda \sum_{k=1}^{M} \gamma_k \sqrt{\sum_{m=1}^{M} \Psi_{mk}^2} \right\}. \tag{4.7}$$

Thus $\widehat{\mathbf{\Psi}}_\lambda \in \mathbb{R}^{M\times M}$ can be interpreted as a Group-Lasso estimator of $\mathbf{\Sigma}$ in the following matrix regression model

$$\widetilde{\mathbf{S}} = \mathbf{\Sigma} + \mathbf{U} + \mathbf{W} \approx \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top + \mathbf{U} + \mathbf{W}, \tag{4.8}$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is a centered error matrix given by $\mathbf{U} = \frac{1}{N}\sum_{i=1}^N \mathbf{U}_i$ and $\mathbf{W} = \frac{1}{N}\sum\limits_{i=1}^N \mathbf{W}_i$. In the above regression model (4.8), there are two errors terms of a different nature. The term $\mathbf{W}$ corresponds to the additive Gaussian errors $\mathcal{E}_1, ..., \mathcal{E}_N$ in model (4.1), while the term $\mathbf{U} = \mathbf{S} - \mathbf{\Sigma}$ represents the difference between the (unobserved) sample covariance matrix $\mathbf{S}$ and the matrix $\mathbf{\Sigma}$ that we want to estimate.

This approach can be interpreted as a thresholding procedure of the entries of an empirical matrix. To see this, consider the simple case where $M = n$ and the basis functions and observations points are chosen such that the matrix $\mathbf{G}$ is orthogonal. Let $\mathbf{Y} = \mathbf{G}^\top\widetilde{\mathbf{S}}\mathbf{G}$ be a transformation of the empirical covariance matrix $\widetilde{\mathbf{S}}$. In the orthogonal case, the folowing proposition shows that the Group-Lasso estimator $\widehat{\mathbf{\Psi}}_\lambda$ defined by (4.7) consists in thresholding the columns/rows of $\mathbf{Y}$ whose ℓ_2-norm is too small, and in multiplying the other columns/rows by weights between 0 and 1. Hence, the Group-Lasso estimate (4.7) can be interpreted as covariance estimation by soft-thresholding of the columns/rows of $\mathbf{Y}$.

Proposition 4.1. *Suppose that $M = n$ and that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$ where $\mathbf{I}_n$ denotes the identity matrix of size $n \times n$. Let $\mathbf{Y} = \mathbf{G}^\top\widetilde{\mathbf{S}}\mathbf{G}$. Then, the Group-Lasso estimator $\widehat{\mathbf{\Psi}}_\lambda$ defined by (4.7) is the $n \times n$ symmetric matrix whose entries are given by*

$$\left(\widehat{\mathbf{\Psi}}_\lambda\right)_{mk} = \begin{cases} 0 & \text{if} \quad \sqrt{\sum_{j=1}^M \mathbf{Y}_{jk}^2} \le \lambda\gamma_k \\ Y_{mk}\left(1 - \frac{\lambda\gamma_k}{\sqrt{\sum_{j=1}^M \mathbf{Y}_{jk}^2}}\right) & \text{if} \quad \sqrt{\sum_{j=1}^M \mathbf{Y}_{jk}^2} > \lambda\gamma_k \end{cases} \tag{4.9}$$

for $1 \le k, m \le M$.

4.3 Consistency of the Group-Lasso estimator

4.3.1 Notations and main assumptions

Let us begin by some definitions. Let β be a vector in $\mathbb{R}^M$. For a subset $J \subset \{1,\ldots,M\}$ of indices of cardinality $|J|$, then β_J is the vector in $\mathbb{R}^M$ that has the same coordinates as β on J and zeros coordinates on the complement J^c of J. The $n \times |J|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in J is denoted by $\mathbf{G}_J$.

The sparsity of $\mathbf{\Psi} \in \mathcal{S}_M$ is defined as its number of non-zero columns (and thus by symmetry non-zero rows) namely

Definition 4.1. *For $\mathbf{\Psi} \in \mathcal{S}_M$, the sparsity of $\mathbf{\Psi}$ is $\mathcal{M}(\mathbf{\Psi}) = |\{k : \mathbf{\Psi}_k \neq \mathbf{0}\}|$.*

Then, let us introduce the following quantities that control the minimal eigenvalues of sub-matrices of small size extracted from the matrix $\mathbf{G}^\top\mathbf{G}$, and the correlations between the columns of $\mathbf{G}$:

Definition 4.2. *Let $0 < s \leq M$. Then,*

$$\rho_{\min}(s) := \inf_{\substack{J \subset \{1, \ldots, M\} \\ |J| \leq s}} \left(\frac{\beta_J^\top \mathbf{G}^\top \mathbf{G} \beta_J}{\|\beta_J\|_{\ell_2}^2} \right) = \inf_{\substack{J \subset \{1, \ldots, M\} \\ |J| \leq s}} \rho_{\min}\left(\mathbf{G}_J^\top \mathbf{G}_J \right),$$

where $\rho_{\min}\left(\mathbf{G}_J^\top \mathbf{G}_J\right)$ denotes the smallest eigenvalue of $\mathbf{G}_J^\top \mathbf{G}_J$.

Definition 4.3. *The mutual coherence $\theta(\mathbf{G})$ of the columns $\mathbf{G}_k$, $k = 1, \ldots, M$ of $\mathbf{G}$ is defined as*

$$\theta(\mathbf{G}) := \max\left\{ \left| \mathbf{G}_{k'}^\top \mathbf{G}_k \right|, \ k \neq k', \ 1 \leq k, k' \leq M \right\},$$

and let

$$\mathbf{G}_{\max}^2 := \max\left\{ \|\mathbf{G}_k\|_{\ell_2}^2, \ 1 \leq k \leq M \right\}.$$

To derive oracle inequalities showing the consistency of the Group-Lasso estimator $\widehat{\mathbf{\Psi}}_\lambda$ the correlations between the columns of $\mathbf{G}$ (measured by $\theta(\mathbf{G})$) should not be too large when compared to the minimal eigenvalues of small matrices extracted from $\mathbf{G}^\top \mathbf{G}$, which is formulated in the following assumption:

Assumption 4.1. *Let $c_0 > 0$ be some constant and $0 < s \leq M$. Then*

$$\theta(\mathbf{G}) < \frac{\rho_{\min}(s)^2}{c_0 \rho_{\max}(\mathbf{G}^\top \mathbf{G}) s},$$

where $\rho_{\max}\left(\mathbf{G}^\top \mathbf{G}\right)$ denotes the largest eigenvalue of $\mathbf{G}^\top \mathbf{G}$.

Assumption 4.1 is inspired by recent results in Bickel, Ritov and Tsybakov [13] on the consistency of Lasso estimators in the standard nonparametric regression model using a large dictionary of basis functions. In Bickel et al. [13], a general condition called *restricted eigenvalue assumption* is introduced to control the minimal eigenvalues of the Gram matrix associated to the dictionary over sets of sparse vectors. In the setting of nonparametric regression, a condition similar to Assumption 4.1 is given in Bickel et al. [13] as an example for which the restricted eigenvalue assumption holds.

Let us give some examples for which Assumption 4.1 is satisfied. If $M \leq n$ and the design points are chosen such that the columns of the matrix $\mathbf{G}$ are orthonormal vectors in $\mathbb{R}^n$, then for any $0 < s \leq M$ one has that $\rho_{\min}(s) = 1$ and $\theta(\mathbf{G}) = 0$ and thus Assumption 4.1 holds for any value of c_0 and s.

Now, suppose that the columns of $\mathbf{G}$ are normalized to one, i.e. $\|\mathbf{G}_k\|_{\ell_2} = 1$ for all $k = 1, \ldots, M$ implying that $\mathbf{G}_{\max} = 1$. Let $\beta \in \mathbb{R}^M$. Then, for any $J \subset \{1, \ldots, M\}$ with $|J| \leq s \leq \min(n, M)$

$$\beta_J^\top \mathbf{G}^\top \mathbf{G} \beta_J \geq \|\beta_J\|_{\ell_2}^2 - \theta(\mathbf{G}) s \|\beta_J\|_{\ell_2}^2,$$

which implies that

$$\rho_{\min}(s) \geq 1 - \theta(\mathbf{G}) s.$$

Therefore, if $(1 - \theta(\mathbf{G})s)^2 > c_0 \theta(\mathbf{G}) \rho_{\max}(\mathbf{G}^\top \mathbf{G}) s$, then Assumption 4.1 is satisfied.

Let us now specify the law of the stochastic process X. For this, recall that for a real-valued random variable Z, the ψ_α Orlicz norm of Z is

$$\|Z\|_{\psi_\alpha} := \inf\left\{C > 0\ ;\ \mathbb{E}\exp\left(\frac{|Z|^\alpha}{C^\alpha}\right) \leq 2\right\}.$$

Such Orlicz norms are useful to characterize the tail behavior of random variables. Indeed, if $\|Z\|_{\psi_\alpha} < +\infty$ then this is equivalent to the statement that there exists two constants $K_1, K_2 > 0$ such that for all $x > 0$ (see e.g. Mendelson and Pajor [69] for more details on Orlicz norms of random variables)

$$\mathbb{P}\left(|Z| \geq x\right) \leq K_1 \exp\left(-\frac{x^\alpha}{K_2^\alpha}\right).$$

Thus, if $\|Z\|_{\psi_2} < +\infty$ then Z is said to have a sub-Gaussian behavior and if $\|Z\|_{\psi_1} < +\infty$ then Z is said to have a sub-Exponential behavior. In the next sections, oracle inequalities for the Group-Lasso estimator will be derived under the following assumption on X:

Assumption 4.2. *The random vector $\mathbf{X} = \left(X\left(t_1\right), ..., X\left(t_n\right)\right)^\top \in \mathbb{R}^n$ is such that*

(A1) *There exists $\rho\left(\mathbf{\Sigma}\right) > 0$ such that, for all vector $\beta \in \mathbb{R}^n$ with $\|\beta\|_{\ell_2} = 1$, then $\left(\mathbb{E}|\mathbf{X}^\top\beta|^4\right)^{1/4} < \rho\left(\mathbf{\Sigma}\right)$.*

(A2) *Set $Z = \|\mathbf{X}\|_{\ell_2}$. There exists $\alpha \geq 1$ such that $\|Z\|_{\psi_\alpha} < +\infty$.*

Note that **(A1)** implies that $\|\mathbf{\Sigma}\|_2 \leq \rho^2\left(\mathbf{\Sigma}\right)$. Indeed, one has that

$$\begin{aligned}\|\mathbf{\Sigma}\|_2 = \rho_{\max}(\mathbf{\Sigma}) &= \sup_{\beta\in\mathbb{R}^n,\ \|\beta\|_{\ell_2}=1} \beta^\top\mathbf{\Sigma}\beta = \sup_{\beta\in\mathbb{R}^n,\ \|\beta\|_{\ell_2}=1} \mathbb{E}\left(\beta^\top\mathbf{X}\mathbf{X}^\top\beta\right) \\ &= \sup_{\beta\in\mathbb{R}^n,\ \|\beta\|_{\ell_2}=1} \mathbb{E}|\beta^\top\mathbf{X}|^2 \leq \sup_{\beta\in\mathbb{R}^n,\ \|\beta\|_{\ell_2}=1} \sqrt{\mathbb{E}|\beta^\top\mathbf{X}|^4} \leq \rho^2\left(\mathbf{\Sigma}\right).\end{aligned}$$

Assumption **(A2)** requires that $\|Z\|_{\psi_\alpha} < +\infty$, where $Z = \|\mathbf{X}\|_{\ell_2}$. The following proposition provides some examples where such an assumption holds.

Proposition 4.2. *Let $Z = \|\mathbf{X}\|_{\ell_2} = \left(\sum_{i=1}^n |X(t_i)|^2\right)^{1/2}$. Then*

- *If X is a Gaussian process*

$$\|Z\|_{\psi_2} < \sqrt{8/3}\sqrt{\mathrm{Tr}(\mathbf{\Sigma})}.$$

- *If the random process X is such that $\|Z\|_{\psi_2} < +\infty$, and there exists a constant C_1 such that $\|\mathbf{\Sigma}_{ii}^{-1/2}|X(t_i)|\|_{\psi_2} \leq C_1$ for all $i = 1, \ldots, n$, then*

$$\|Z\|_{\psi_2} < C_1\sqrt{\mathrm{Tr}(\mathbf{\Sigma})}.$$

- *If X is a bounded process, meaning that there exists a constant $R > 0$ such that for all $t \in T$, $|X(t)| \leq R$, then for any $\alpha \geq 1$,*

$$\|Z\|_{\psi_\alpha} \leq \sqrt{n}R(\log 2)^{-1/\alpha}.$$

Assumption 4.2 will be used to control the deviation in operator norm between the sample covariance matrix $\mathbf{S}$ and the true covariance matrix $\mathbf{\Sigma}$ in the sense of the following proposition whose proof follows from Theorem 2.1 in Mendelson and Pajor [69].

Proposition 4.3. *Let $X_1, ..., X_N$ be independent copies of the stochastic process X, let $Z = \|\mathbf{X}\|_{\ell_2}$ and $\mathbf{X}_i = (X_i(t_1), ..., X_i(t_n))^\top$ for $i = 1, \dots, N$. Recall that $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^\top$ and $\mathbf{\Sigma} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$. Suppose that X satisfies Assumption 4.2. Let $d = \min(n, N)$. Then, there exists a universal constant $\delta_* > 0$ such that for all $x > 0$,*

$$\mathbb{P}\left(\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_2 \geqslant \tau_{d,N,n} x\right) \leqslant \exp\left(-(\delta_*^{-1} x)^{\frac{\alpha}{2+\alpha}}\right), \tag{4.10}$$

where $\tau_{d,N,n} = \max(A_{d,N}^2, B_{d,N})$, with

$$A_{d,N} = \|Z\|_{\psi_\alpha} \frac{\sqrt{\log d}(\log N)^{1/\alpha}}{\sqrt{N}} \text{ and } B_{d,N} = \frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}} + \|\mathbf{\Sigma}\|_2^{1/2} A_{d,N}.$$

Let us briefly comment Proposition 4.3 in some specific cases. If X is Gaussian, then Proposition 4.2 implies that $A_{d,N} \leq A_{d,N,1}$ where

$$A_{d,N,1} = \sqrt{8/3}\sqrt{\mathrm{Tr}(\mathbf{\Sigma})}\frac{\sqrt{\log d}(\log N)^{1/\alpha}}{\sqrt{N}} \leq \sqrt{8/3}\,\|\mathbf{\Sigma}\|_2^{1/2}\sqrt{\frac{n}{N}}\sqrt{\log d}(\log N)^{1/\alpha}, \tag{4.11}$$

and in this case inequality (4.10) becomes

$$\mathbb{P}\left(\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_2 \geqslant \max\left(A_{d,N,1}^2, B_{d,N,1}\right) x\right) \leqslant \exp\left(-(\delta_*^{-1} x)^{\frac{\alpha}{2+\alpha}}\right) \tag{4.12}$$

for all $x > 0$, where $B_{d,N,1} = \frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}} + \|\mathbf{\Sigma}\|_2^{1/2} A_{d,N,1}$.

If X is a bounded process by some constant $R > 0$, then using Proposition 4.2 and by letting $\alpha \to +\infty$, Proposition 4.3 implies that for all $x > 0$

$$\mathbb{P}\left(\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_2 \geqslant \max\left(A_{d,N,2}^2, B_{d,N,2}\right) x\right) \leqslant \exp\left(-\delta_*^{-1} x\right), \tag{4.13}$$

where

$$A_{d,N,2} = R\sqrt{\frac{n}{N}}\sqrt{\log d} \text{ and } B_{d,N,2} = \frac{\rho^2(\mathbf{\Sigma})}{\sqrt{N}} + \|\mathbf{\Sigma}\|_2^{1/2} A_{d,N,2}. \tag{4.14}$$

Inequalities (4.12) and (4.13) illustrate the fact that in a high-dimensional setting when $n >> N$ or when n and N are of the same magnitude ($\frac{n}{N} \to c > 0$ as $n, N \to +\infty$) then $\|\mathbf{S} - \mathbf{\Sigma}\|_2$ does not converge to zero (in probability). Therefore, without any further restrictions on the structure of the covariance matrix $\mathbf{\Sigma}$, then $\mathbf{S}$ is not a consistent estimator.

Now, let us explain how supposing that X has a sparse representation in a dictionary of basis functions may improve the quality of $\mathbf{S}$ as an estimator of $\mathbf{\Sigma}$. To see this, consider the simplest case $X = X^0$, where the process X^0 has a sparse representation in the basis $(g_m)_{1 \leq m \leq M}$ given by

$$X^0(t) = \sum_{m \in J^*} a_m g_m(t),\ t \in T, \tag{4.15}$$

where $J^* \subset \{1, \ldots, M\}$ is a subset of indices of cardinality $|J^*| = s^*$ and a_m, $m \in J^*$ are random coefficients (possibly correlated). Under such an assumption, the following proposition holds.

Proposition 4.4. *Suppose that $X = X^0$ with X^0 defined by (4.15) with $s^* \leq \min(n, M)$. Assume that X satisfies Assumption 4.2 and that the matrix $\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}$ is invertible, where $\mathbf{G}_{J^*}$ denotes the $n \times |J^*|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in J^*. Then, there exists a universal constant $\delta_* > 0$ such that for all $x > 0$,*

$$\mathbb{P}\left(\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_2 \geqslant \widetilde{\tau}_{d^*,N,s^*} x\right) \leqslant \exp\left(-(\delta_*^{-1} x)^{\frac{\alpha}{2+\alpha}}\right), \tag{4.16}$$

where $\widetilde{\tau}_{d^*,N,s^*} = \max(\widetilde{A}^2_{d^*,N,s^*}, \widetilde{B}_{d^*,N,s^*})$, with

$$\widetilde{A}_{d^*,N,s^*} = \rho_{\max}^{1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \|\widetilde{Z}\|_{\psi_\alpha} \frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$$

and

$$\widetilde{B}_{d^*,N,s^*} = \left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right) \frac{\rho^2\left(\mathbf{\Sigma}\right)}{\sqrt{N}} + \left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right)^{1/2} \|\mathbf{\Sigma}\|_2^{1/2} \widetilde{A}_{d^*,N,s^*},$$

with $d^* = \min(N, s^*)$ and $\widetilde{Z} = \|\mathbf{a}_{J^*}\|_{\ell_2}$, where $\mathbf{a}_{J^*} = (\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1} \mathbf{G}_{J^*}^\top \mathbf{X} \in \mathbb{R}^{s^*}$.

Using Proposition 4.2 and Proposition 4.4 it follows that:

- If $X = X^0$ is a Gaussian process then

$$\widetilde{A}_{d^*,N,s^*} \leq \sqrt{8/3}\left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right)^{1/2} \|\mathbf{\Sigma}\|_2^{1/2} \sqrt{\frac{s^*}{N}} \sqrt{\log d^*}(\log N)^{1/\alpha}. \tag{4.17}$$

- If $X = X^0$ is such that the random variables a_m are bounded by for some constant $R > 0$, then

$$\widetilde{A}_{d^*,N,s^*} \leq R\|g\|_\infty \sqrt{\frac{s^*}{N}} \sqrt{\log d^*}, \tag{4.18}$$

  with $\|g\|_\infty = \max_{1 \leq m \leq M} \|g_m\|_\infty$ where $\|g_m\|_\infty = \sup_{t \in \mathcal{T}} |g_m(t)|$.

Therefore, let us compare the bounds (4.17) and (4.18) with the inequalities (4.11) and (4.14). It follows that, in the case $X = X^0$, if the sparsity s^* of X in the dictionary is small compared to the number of time points n then the deviation between $\mathbf{S}$ and $\mathbf{\Sigma}$ is much smaller than in the general case without any assumption on the structure of $\mathbf{\Sigma}$. Obviously, the gain also depends on the control of the ratio $\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}$. Note that in the case of an orthonormal design ($M = n$ and $\mathbf{G}^\top \mathbf{G} = \mathbf{I}_n$) then $\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) = \rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) = 1$ for any J^*, and thus the gain in operator norm between $\mathbf{S}$ and $\mathbf{\Sigma}$ clearly depends on the size of $\frac{s^*}{N}$ compared to $\frac{n}{N}$. Supposing that $X = X^0$ also implies that the operator norm of the error term $\mathbf{U}$ in the matrix regression model (4.8) is controlled by the ratio $\frac{s^*}{N}$ instead of the ratio $\frac{n}{N}$ when no assumptions are made on the structure of $\mathbf{\Sigma}$. This means that if X has a sparse representation in the dictionary then the error term $\mathbf{U}$ becomes smaller.

4.3.2 An oracle inequality for the Frobenius norm

Consistency is first studied for the normalized Frobenius norm $\frac{1}{n}\|\mathbf{A}\|_F^2$ of an $n \times n$ matrix $\mathbf{A}$. The following theorem provides an oracle inequality for the Group-Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda = \mathbf{G}\widehat{\mathbf{\Psi}}_\lambda \mathbf{G}^\top$.

Theorem 4.1. *Assume that X satisfies Assumption 4.2. Let $1 \leq s \leq \min(n, M)$ and $\epsilon > 0$. Suppose that Assumption 4.1 holds with $c_0 = 3 + 4/\epsilon$. Consider the Group-Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ defined by (4.5) with the choices*

$$\gamma_k = 2\|\mathbf{G}_k\|_{\ell_2}\sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)}$$

and

$$\lambda = \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^2 \text{ for some constant } \delta > 1,$$

where $\mathbf{\Sigma}_{noise} = \mathbb{E}(\mathbf{W}_1)$. Then, with probability at least $1 - M^{1-\delta}$ one has that

$$\begin{aligned}\frac{1}{n}\left\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\right\|_F^2 \quad \leq \quad & (1+\epsilon) \inf_{\substack{\mathbf{\Psi} \in \mathcal{S}_M \\ \mathcal{M}(\mathbf{\Psi}) \leq s}} \left(\frac{4}{n}\left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{\Sigma}\right\|_F^2 + \frac{8}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2\right. \\ & \left. + C(\epsilon)\frac{\mathbf{G}^2_{\max}\rho_{\max}(\mathbf{G}^\top\mathbf{G})}{\kappa^2_{s,c_0}}\lambda^2\frac{\mathcal{M}(\mathbf{\Psi})}{n}\right), \end{aligned} \tag{4.19}$$

where $\kappa^2_{s,c_0} = \rho_{\min}(s)^2 - c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})s$, and $C(\epsilon) = 8\frac{\epsilon}{1+\epsilon}(1 + 2/\epsilon)^2$.

The first term $\frac{1}{n}\left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{\Sigma}\right\|_F^2$ in inequality (4.19) is the bias of the estimator $\widehat{\mathbf{\Sigma}}_\lambda$. It reflects the quality of the approximation of $\mathbf{\Sigma}$ by the set of matrices of the form $\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top$, with $\mathbf{\Psi} \in \mathcal{S}_M$ and $\mathcal{M}(\mathbf{\Psi}) \leq s$. As an example, suppose that $X = X^0$, where the process X^0 has a sparse representation in the basis $(g_m)_{1 \leq m \leq M}$ given by

$$X^0(t) = \sum_{m \in J^*} a_m g_m(t), \; t \in T,$$

where $J^* \subset \{1, \ldots, M\}$ is a subset of indices of cardinality $|J^*| = s^* \leq s$ and $a_m, m \in J^*$ are random coefficients. Then, in this case, since $s^* \leq s$ the bias term in (4.19) is equal to zero.

The second term $\frac{1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2$ in (4.19) is a variance term as the empirical covariance matrix $\mathbf{S}$ is an unbiased estimator of $\mathbf{\Sigma}$. Using the inequality $\frac{1}{n}\|\mathbf{A}\|_F^2 \leq \|\mathbf{A}\|_2^2$ that holds for any $n \times n$ matrix $\mathbf{A}$, it follows that $\frac{1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 \leq \|\mathbf{S} - \mathbf{\Sigma}\|_2^2$. Therefore, under the assumption that X has a sparse representation in the dictionary (e.g. when $X = X_0$ as above) then the variance term $\frac{1}{n}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2$ is controlled by the ratio $\frac{s^*}{N} \leq \frac{s}{N}$ (see Proposition 4.4) instead of the ratio $\frac{n}{N}$ without any assumption on the structure of $\mathbf{\Sigma}$.

The third term in (4.19) is also a variance term due to the noise in the measurements (4.1). If there exists a constant $c > 0$ independent of n and N such that $\frac{n}{N} \leq c$ then the decay of this third variance term is essentially controlled by the ratio $\frac{\mathcal{M}(\mathbf{\Psi})}{n} \leq \frac{s}{n}$.

Therefore, if $\mathcal{M}(\boldsymbol{\Psi}) \leq s$ with sparsity s much smaller than n then the variance of the Group-Lasso estimator $\widehat{\boldsymbol{\Sigma}}_\lambda$ is smaller than the variance of $\widetilde{\mathbf{S}}$. This shows some of the improvements achieved by regularization (4.7) of the empirical covariance matrix $\widetilde{\mathbf{S}}$ with a Group-Lasso penalty.

4.3.3 An oracle inequality for the operator norm

The "normalized" Frobenius norm $\frac{1}{n}\left\|\widehat{\boldsymbol{\Sigma}}_\lambda - \boldsymbol{\Sigma}\right\|_F^2$ (the average of the eigenvalues) can be viewed as a reasonable proxy for the operator norm $\left\|\widehat{\boldsymbol{\Sigma}}_\lambda - \boldsymbol{\Sigma}\right\|_2^2$. It is thus expected that the results of Theorem 4.1 imply that the Group-Lasso estimator $\widehat{\boldsymbol{\Sigma}}_\lambda$ is a good estimator of $\boldsymbol{\Sigma}$ in operator norm. To see this, let us consider the case where $\widetilde{X}$ consists in noisy observations of the process X^0 (4.15), meaning that

$$\widetilde{X}(t_j) = X^0(t_j) + \mathcal{E}(t_j), \; j = 1, \ldots, n, \tag{4.20}$$

where $\mathcal{E}$ is a second order Gaussian process with zero mean independent of X^0. In this case, one has that

$$\boldsymbol{\Sigma} = \mathbf{G}\boldsymbol{\Psi}^*\mathbf{G}^\top, \text{ where } \boldsymbol{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right),$$

where $\mathbf{a}$ is the random vector of $\mathbb{R}^M$ with $\mathbf{a}_m = a_m$ for $m \in J^*$ and $\mathbf{a}_m = 0$ for $m \notin J^*$. Therefore, using Theorem 4.1 with $s = |J^*| = s^*$, since $\boldsymbol{\Psi}^* \in \{\boldsymbol{\Psi} \in \mathcal{S}_M : M(\boldsymbol{\Psi}) \leq s^*\}$, one can derive the following corollary:

Corollary 4.1. *Suppose that the observations are i.i.d. random variables from model (4.20) and that the conditions of Theorem 4.1 are satisfied with* $1 \leq s = s^* \leq \min(n, M)$. *Then, with probability at least* $1 - M^{1-\delta}$ *one has that*

$$\frac{1}{n}\left\|\widehat{\boldsymbol{\Sigma}}_\lambda - \boldsymbol{\Sigma}\right\|_F^2 \leq C_0\left(n, M, N, s^*, \mathbf{S}, \boldsymbol{\Psi}^*, \mathbf{G}, \boldsymbol{\Sigma}_{noise}\right), \tag{4.21}$$

where

$$C_0\left(n, M, N, s^*, \mathbf{S}, \boldsymbol{\Psi}^*, \mathbf{G}, \boldsymbol{\Sigma}_{noise}\right) = (1+\epsilon)\left(\frac{8}{n}\left\|\mathbf{S} - \mathbf{G}\boldsymbol{\Psi}^*\mathbf{G}^\top\right\|_F^2 + C(\epsilon)\frac{\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G})}{\kappa_{s^*,c_0}^2}\lambda^2\frac{s^*}{n}\right).$$

To simplify notations, write $\widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{\Psi}}_\lambda$, with $\widehat{\boldsymbol{\Psi}}_\lambda$ given by (4.7). Define $\widehat{J}_\lambda \subset \{1, \ldots, M\}$ as

$$\widehat{J}_\lambda \equiv \widehat{J} := \left\{k : \frac{\delta_k}{\sqrt{n}}\left\|\widehat{\boldsymbol{\Psi}}_k\right\|_{\ell_2} > C_1\left(n, M, N, s^*, \mathbf{S}, \boldsymbol{\Psi}^*, \mathbf{G}, \boldsymbol{\Sigma}_{noise}\right)\right\}, \text{ with } \delta_k = \frac{\|\mathbf{G}_k\|_{\ell_2}}{\mathbf{G}_{\max}}, \tag{4.22}$$

and

$$C_1\left(n, M, N, s^*, \mathbf{S}, \boldsymbol{\Psi}^*, \mathbf{G}, \boldsymbol{\Sigma}_{noise}\right) = \frac{4(1+\epsilon)\sqrt{s^*}}{\epsilon\kappa_{s^*,c_0}}\sqrt{C_0\left(n, M, N, s^*, \mathbf{S}, \boldsymbol{\Psi}^*, \mathbf{G}, \boldsymbol{\Sigma}_{noise}\right)}. \tag{4.23}$$

The set of indices $\widehat{J}$ is an estimation of the set of active basis functions J^*. Note that such thresholding procedure (4.22) does not lead immediately to a practical way to choose the set $\widehat{J}$. Indeed the constant C_1 in (4.22) depends on the a priori unknown sparsity s^* and on the amplitude of the noise in the matrix regression model (4.8) measured by the quantities $\frac{8}{n}\left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top\right\|_F^2$ and $\|\mathbf{\Sigma}_{noise}\|_2^2$. Nevertheless, in Section 4.4 on numerical experiments we give a simple procedure to automatically threshold the ℓ_2-norm of columns of the matrix $\widehat{\Psi}_\lambda$ that are two small.

Note that to estimate J^* we did not simply take $\widehat{J} = \widehat{J}_0 := \left\{k : \left\|\widehat{\mathbf{\Psi}}_k\right\|_{\ell_2} \neq 0\right\}$, but rather apply a thresholding step to discard the columns of $\widehat{\mathbf{\Psi}}$ whose ℓ_2-norm is too small. By doing so, we want to stress the fact that to obtain a consistent procedure in operator norm it is not sufficient to simply take $\widehat{J} = \widehat{J}_0$. A similar thresholding step is proposed in Lounici [61] and Lounici, Pontil, Tsybakov and van de Geer [62] in the standard linear model to select a sparse set of active variables when using regularization by a Lasso penalty. In the paper ([61]), the second thresholding step used to estimate the true sparsity pattern depends on a constant (denoted by r) that is related to the amplitude of the unknow coefficients to estimate. Then, the following theorem holds.

Theorem 4.2. *Under the assumptions of Corollary 4.1, for any solution of problem (4.7), we have that with probability at least* $1 - M^{1-\delta}$

$$\max_{1\leq k\leq M} \frac{\delta_k}{\sqrt{n}} \left\|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^*\right\|_{\ell_2} \leq C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right). \tag{4.24}$$

If in addition

$$\min_{k\in J^*} \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2} > 2C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right) \tag{4.25}$$

then with the same probability the set of indices $\widehat{J}$, defined by (4.22), estimates correctly the true set of active basis functions J^*, that is $\widehat{J} = J^*$ with probability at least $1 - M^{1-\delta}$.

The results of Theorem 4.2 indicate that if the ℓ_2-norm of the columns of $\mathbf{\Psi}_k^*$ for $k \in J^*$ are sufficiently large with respect to the level of noise in the matrix regression model (4.8) and the sparsity s^*, then $\widehat{J}$ is a consistent estimation of the active set of variables. Indeed, if $\mathcal{M}\left(\mathbf{\Psi}^*\right) = s^*$, then by symmetry the columns of $\mathbf{\Psi}^*$ such $\mathbf{\Psi}_k^* \neq 0$ have exactly s^* non-zero entries. Hence, the condition (4.25) means that the ℓ_2-norm of $\mathbf{\Psi}_k^* \neq 0$ (normalized by $\frac{\delta_k}{\sqrt{n}}$) has to be larger than $\frac{4(1+\epsilon)}{\epsilon\kappa_{s^*,c_0}}\sqrt{s^*}\sqrt{C_0}$. A simple condition to satisfy such an assumption is that the amplitude of the s^* non-vanishing entries of $\mathbf{\Psi}_k^* \neq 0$ are larger than $\frac{\sqrt{n}}{\delta_k}\frac{4(1+\epsilon)}{\epsilon\kappa_{s^*,c_0}}\sqrt{C_0}$ which can be interpreted as a kind of measure of the noise in model (4.8). This suggests to take as a final estimator of $\mathbf{\Sigma}$ the following matrix:

$$\widehat{\mathbf{\Sigma}}_{\widehat{J}} = \mathbf{G}_{\widehat{J}}\widehat{\mathbf{\Psi}}_{\widehat{J}}\mathbf{G}_{\widehat{J}}, \tag{4.26}$$

where $\mathbf{G}_{\widehat{J}}$ denotes the $n \times |\widehat{J}|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in $\widehat{J}$, and

$$\widehat{\mathbf{\Psi}}_{\widehat{J}} = \underset{\mathbf{\Psi}\in\mathcal{S}_{|\widehat{J}|}}{\operatorname{argmin}} \left\{\left\|\widetilde{\mathbf{S}} - \mathbf{G}_{\widehat{J}}\mathbf{\Psi}\mathbf{G}_{\widehat{J}}^\top\right\|_F^2\right\},$$

where $\mathcal{S}_{|\widehat{J}|}$ denotes the set of $|\widehat{J}|\times|\widehat{J}|$ symmetric matrices. Note that if $\mathbf{G}_{\widehat{J}}^{\top}\mathbf{G}_{\widehat{J}}$ is invertible, then

$$\widehat{\mathbf{\Psi}}_{\widehat{J}} = \left(\mathbf{G}_{\widehat{J}}^{\top}\mathbf{G}_{\widehat{J}}\right)^{-1}\mathbf{G}_{\widehat{J}}^{\top}\widetilde{\mathbf{S}}\mathbf{G}_{\widehat{J}}\left(\mathbf{G}_{\widehat{J}}^{\top}\mathbf{G}_{\widehat{J}}\right)^{-1}.$$

Let us recall that if the observations are i.i.d. random variables from model (4.20) then

$$\mathbf{\Sigma} = \mathbf{G}\mathbf{\Psi}^{*}\mathbf{G}^{\top},$$

where $\mathbf{\Psi}^{*} = \mathbb{E}\left(\mathbf{a}\mathbf{a}^{\top}\right)$ and $\mathbf{a}$ is the random vector of $\mathbb{R}^{M}$ with $\mathbf{a}_{m} = a_{m}$ for $m \in J^{*}$ and $\mathbf{a}_{m} = 0$ for $m \notin J^{*}$. Then, define the random vector $\mathbf{a}_{J^{*}} \in \mathbb{R}^{J^{*}}$ whose coordinates are the random coefficients a_{m} for $m \in J^{*}$. Let $\mathbf{\Psi}_{J^{*}} = \mathbb{E}\left(\mathbf{a}_{J^{*}}\mathbf{a}_{J^{*}}^{\top}\right)$ and denote by $\mathbf{G}_{J^{*}}$ the $n\times|J^{*}|$ matrix obtained by removing the columns of $\mathbf{G}$ whose indices are not in J^{*}. Note that $\mathbf{\Sigma} = \mathbf{G}_{J^{*}}\mathbf{\Psi}_{J^{*}}\mathbf{G}_{J^{*}}^{\top}$.

Assuming that $\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}$ is invertible, define the matrix

$$\mathbf{\Sigma}_{J^{*}} = \mathbf{\Sigma} + \mathbf{G}_{J^{*}}(\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}})^{-1}\mathbf{G}_{J^{*}}^{\top}\mathbf{\Sigma}_{noise}\mathbf{G}_{J^{*}}\left(\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}\right)^{-1}\mathbf{G}_{J^{*}}^{\top}. \tag{4.27}$$

Then, the following theorem gives a control of deviation between $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ and $\mathbf{\Sigma}_{J^{*}}$ in operator norm.

Theorem 4.3. *Suppose that the observations are i.i.d. random variables from model (4.20) and that the conditions of Theorem 4.1 are satisfied with $1 \leq s = s^{*} \leq \min(n, M)$. Suppose that $\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}$ is an invertible matrix, and that*

$$\min_{k\in J^{*}}\frac{\delta_{k}}{\sqrt{n}}\|\mathbf{\Psi}_{k}^{*}\|_{\ell_{2}} > 2C_{1}\left(n, M, N, s^{*}, \mathbf{S}, \mathbf{\Psi}^{*}, \mathbf{G}, \mathbf{\Sigma}_{noise}\right),$$

where $C_{1}\left(n, M, N, s^{*}, \mathbf{S}, \mathbf{\Psi}^{*}, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)$ is the constant defined in (4.23).

Let $\mathbf{Y} = \left(\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}\right)^{-1}\mathbf{G}_{J^{*}}^{\top}\widetilde{\mathbf{X}}$ and $\widetilde{Z} = \|\mathbf{Y}\|_{\ell_{2}}$. Let $\rho^{4}\left(\mathbf{\Sigma}_{noise}\right) = \sup\limits_{\beta\in\mathbb{R}^{n},\|\beta\|_{\ell_{2}}=1}\mathbb{E}|\mathcal{E}^{\top}\beta|^{4}$, where $\mathcal{E} = \left(\mathcal{E}\left(t_{1}\right), ..., \mathcal{E}\left(t_{n}\right)\right)^{\top}$. Then, with probability at least $1 - M^{1-\delta} - M^{-\left(\frac{\delta_{\star}}{\delta_{*}}\right)^{\frac{\alpha}{2+\alpha}}}$, with $\delta > 1$ and $\delta_{\star} > \delta_{*}$ one has that

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^{*}}\right\|_{2} \leq \rho_{\max}\left(\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}\right)\widetilde{\tau}_{N,s^{*}}\delta_{\star}\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}, \tag{4.28}$$

where $\widetilde{\tau}_{N,s^{*}} = \max(\widetilde{A}_{N,s^{*}}^{2}, \widetilde{B}_{N,s^{*}})$ with

$$\widetilde{A}_{N,s^{*}} = \|\widetilde{Z}\|_{\psi_{\alpha}}\frac{\sqrt{\log d^{*}}(\log N)^{1/\alpha}}{\sqrt{N}}$$

and

$$\widetilde{B}_{N,s^{*}} = \frac{\widetilde{\rho}^{2}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})\rho_{\min}^{-1}\left(\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}\right)}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^{*}}\|_{2} + \rho_{\min}^{-1}\left(\mathbf{G}_{J^{*}}^{\top}\mathbf{G}_{J^{*}}\right)\|\mathbf{\Sigma}_{noise}\|_{2}\right)^{1/2}\widetilde{A}_{N,s^{*}},$$

where $d^{*} = \min(N, s^{*})$ and $\widetilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) = 8^{1/4}\left(\rho^{4}\left(\mathbf{\Sigma}\right) + \rho^{4}\left(\mathbf{\Sigma}_{noise}\right)\right)^{1/4}$.

First note that the above theorem gives a deviation in operator norm from $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ to the matrix $\mathbf{\Sigma}_{J^*}$ (4.27) which is not equal to the true covariance $\mathbf{\Sigma}$ of X at the design points. Indeed, even if we know the true sparsity set J^*, the additive noise in the measurements in model (4.1) complicates the estimation of $\mathbf{\Sigma}$ in operator norm. However, although $\mathbf{\Sigma}_{J^*} \neq \mathbf{\Sigma}$, they can have the same eigenvectors if the structure of the additive noise matrix term in (4.27) is not too complex. As an example, consider the case of an additive white noise, for which $\mathbf{\Sigma}_{noise} = \sigma^2 \mathbf{I}_n$, where σ is the level of noise and $\mathbf{I}_n$ the $n \times n$ identity matrix. Under such an assumption, if we further suppose for simplicity that $(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1} = \mathbf{I}_{s^*}$, then $\mathbf{\Sigma}_{J^*} = \mathbf{\Sigma} + \sigma^2 \mathbf{G}_{J^*}(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1}\mathbf{G}_{J^*}^\top = \mathbf{\Sigma} + \sigma^2 \mathbf{I}_n$ and clearly $\mathbf{\Sigma}_{J^*}$ and $\mathbf{\Sigma}$ have the same eigenvectors. Therefore, the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ can be used as estimators of the eigenvectors of $\mathbf{\Sigma}$ which is suitable for the sparse PCA application described in the next section on numerical experiments.

Let us illustrate the implications of Theorem 4.3 on a simple example. If X is Gaussian, the random vector $\mathbf{Y} = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top \left(\mathbf{X} + \mathcal{E}\right)$ is also Gaussian and Proposition 4.2 can be used to prove that

$$\begin{aligned}
\|\widetilde{Z}\|_{\psi_2} &\leq \sqrt{8/3}\sqrt{\operatorname{Tr}\left(\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top \left(\mathbf{\Sigma} + \mathbf{\Sigma}_{noise}\right) \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}\right)} \\
&\leq \sqrt{8/3}\|\mathbf{\Sigma} + \mathbf{\Sigma}_{noise}\|_2^{1/2} \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\sqrt{s^*}
\end{aligned}$$

and Theorem 4.3 implies that with high probability

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \widetilde{\tau}_{N,s^*,1}\delta\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}},$$

where $\widetilde{\tau}_{N,s^*,1} = \max(\widetilde{A}^2_{N,s^*,1}, \widetilde{B}_{N,s^*,1})$, with

$$\widetilde{A}_{N,s^*,1} = \sqrt{8/3}\|\mathbf{\Sigma} + \mathbf{\Sigma}_{noise}\|_2^{1/2}\rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\sqrt{\log d^*}(\log N)^{1/\alpha}\sqrt{\frac{s^*}{N}}$$

and

$$\widetilde{B}_{N,s^*,1} = \frac{\widetilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})\rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2}\widetilde{A}_{N,s^*,1}.$$

Therefore, in the Gaussian case (but also under other assumptions for X such as those in Proposition 4.2) the above equations show that the operator norm $\left\|\widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2^2$ depends on the ratio $\frac{s^*}{N}$. Recall that $\|\mathbf{S} - \mathbf{\Sigma}\|_2^2$ depends on the ratio $\frac{n}{N}$. Thus, using $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ clearly yields significant improvements if s^* is small compared to n.

To summarize our results let us finally consider the case of an orthogonal design. Combining Theorems 4.1, 4.2 and 4.3 one arrives at the following corrolary:

Corollary 4.2. *Suppose that the observations are i.i.d. random variables from model (4.20). Suppose that $M = n$ and that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$ (orthogonal design) and that X^0 satisfies Assumption 4.2. Let $\epsilon > 0$ and $1 \leq s^* \leq \min(n, M)$. Consider the Group-Lasso estimator $\widehat{\mathbf{\Sigma}}_\lambda$ defined by (4.5) with the choices*

$$\gamma_k = 2, k = 1, \ldots, n \text{ and } \lambda = \|\mathbf{\Sigma}_{noise}\|_2\left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta\log M}{N}}\right)^2 \text{ for } \delta > 1.$$

Suppose that

$$\min_{k \in J^*} \|\mathbf{\Psi}_k^*\|_{\ell_2} > 2n^{1/2}\widetilde{C}_1\left(\sigma, n, s^*, N, \delta\right), \tag{4.29}$$

where $\widetilde{C}_1\left(\sigma, n, s, N, \delta\right) = \frac{4(1+\epsilon)\sqrt{s^*}}{\epsilon}\sqrt{\widetilde{C}_0\left(\sigma, n, s^*, N, \delta\right)}$ *and*

$$\widetilde{C}_0\left(\sigma, n, s^*, N, \delta\right) = (1+\epsilon)\left(\frac{8}{n}\left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}^*\mathbf{G}^\top\right\|_F^2 + C(\epsilon)\lambda^2\frac{s^*}{n}\right).$$

Take $\widehat{J} := \left\{k : \left\|\widehat{\mathbf{\Psi}}_k\right\|_{\ell_2} > n^{1/2}\widetilde{C}_1\left(\sigma, n, s, N, \delta\right)\right\}$. *Let* $\mathbf{Y} = \mathbf{G}_{J^*}^\top\widetilde{\mathbf{X}}$ *and* $\widetilde{Z} = \|\mathbf{Y}\|_{\ell_2}$. *Then, with probability at least* $1 - M^{1-\delta} - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$, *with* $\delta > 1$ *and* $\delta_\star > \delta_*$ *one has that*

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \widetilde{\tau}_{N,s^*}\delta_\star\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}, \tag{4.30}$$

where $\widetilde{\tau}_{N,s^*} = \max(\widetilde{A}_{N,s^*}^2, \widetilde{B}_{N,s^*})$, *with*

$$\widetilde{A}_{N,s^*} = \|\widetilde{Z}\|_{\psi_\alpha}\frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$$

and

$$\widetilde{B}_{N,s^*} = \frac{\widetilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2}\widetilde{A}_{N,s^*}.$$

4.4 Numerical experiments and an application to sparse PCA

In this section we present some simulated examples to illustrate the practical behaviour of the covariance matrix estimator by group Lasso regularization proposed in this paper. In particular, we show its performances with an application to sparse Principal Components Analysis (PCA). In the numerical experiments, we use the explicit estimator described in Proposition 4.1 in the case $M = n$ and an orthogonal design matrix $\mathbf{G}$, and also the estimator proposed in the more general situation when $n < M$. The programs for our simulations were implemented using the MATLAB programming environment.

4.4.1 Description of the estimating procedure and the data

We consider a noisy stochastic processes $\widetilde{X}$ on $\mathbb{T} = [0, 1]$ with values in $\mathbb{R}$ observed at fixed location points $t_1, ..., t_n$ in $[0, 1]$, generated according to

$$\widetilde{X}(t_j) = X^0(t_j) + \sigma\epsilon_j, \ j = 1, \ldots, n, \tag{4.31}$$

where $\sigma > 0$ is the level of noise, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. standard Gaussian variables, and X^0 is a random process independent of the ϵ_j's given by

$$X^0(t) = af(t),$$

where a is a Gaussian random coefficient such that $\mathbb{E}a = 0$, $\mathbb{E}a^2 = \gamma^2$, and $f : [0,1] \rightarrow \mathbb{R}$ is an unknown function. The simulated data consists in a sample of N independent observations of the process $\widetilde{X}$ at the points $t_1, ..., t_n$, which are generated according to (4.31). Note that in this setting $\mathbf{\Sigma}_{noise} = \sigma^2 \mathbf{I}_n$.

The covariance matrix $\mathbf{\Sigma}$ of the process X^0 at the locations points $t_1, ..., t_n$ is given by $\mathbf{\Sigma} = \gamma^2 \mathbf{F}\mathbf{F}^\top$, where $\mathbf{F} = (f(t_1), ..., f(t_1))^\top \in \mathbb{R}^n$. Note that the largest eigenvalue of $\mathbf{\Sigma}$ is $\gamma^2 \|\mathbf{F}\|_{\ell_2}^2$ with corresponding eigenvector $\mathbf{F}$.

We suppose that the signal f has some sparse representation in a large dictionary of basis functions of size M, given by $\{g_m, m = 1, \ldots, M\}$, meaning that $f(t) = \sum_{m=1}^{M} \beta_m g_m(t)$, with $J^* = \{m, \beta_m \neq 0\}$ of small cardinality s^*. Then, the process X^0 can be written as $X^0(t) = \sum_{m=1}^{M} a\beta_m g_m(t)$, and thus $\mathbf{\Sigma} = \gamma^2 \mathbf{G}\mathbf{\Psi}_{J^*}\mathbf{G}^\top$, where $\mathbf{\Psi}_{J^*}$ is an $M \times M$ matrix with entries equal to $\beta_m \beta_{m'}$ for $1 \leq m, m' \leq M$. Note that the number of non-vanishing columns of $\mathbf{\Psi}_{J^*}$ is $\mathcal{M}(\mathbf{\Psi}_{J^*}) = s^*$ and $\mathbf{F} = \mathbf{G}\beta$.

We aim at estimating $\mathbf{F}$ by the eigenvector corresponding to the largest eigenvalue of the matrix $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ defined in (4.26), in a high-dimensional setting with $n > N$. The idea behind this is that $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ is a consistent estimator of $\mathbf{\Sigma}_{J^*}$ (see its definition in 4.27) in operator norm, and thus the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ can be used as estimators of the eigenvectors of $\mathbf{\Sigma}_{J^*}$. Although the matrices $\mathbf{\Sigma}_{J^*}$ and $\mathbf{\Sigma}$ may have different eigenvectors, we will show in the examples below that depending on the design points and for appropriate values of the level of noise σ, the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ can be used as estimators of the eigenvectors of $\mathbf{\Sigma}$.

The estimator $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ of the covariance matrix $\mathbf{\Sigma}$ is computed as follows. First, the size of the dictionary M is specified, as well as the basis functions $\{g_m, m = 1, ..., M\}$. In the examples below, we will use for the test function f the signals HeaviSine and Blocks (see e.g. Antoniadis, Bigot and Sapatinas [4] for a definition), and the Symmlet 8 and Haar wavelet basis (for HeaviSine and Blocks respectively), which are implemented in the Matlab's open-source library WaveLab (see e.g. Antoniadis et al. [4] for further references on wavelet methods in nonparametric statistics).

Then, we compute the covariance group Lasso (CGL) estimator $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}} = \mathbf{G}\widehat{\mathbf{\Psi}}_{\widehat{\lambda}}\mathbf{G}^\top$, where $\widehat{\mathbf{\Psi}}_{\widehat{\lambda}}$ is defined in (4.7). We use a completely data-driven choice for the regularizarion parameter λ, given by $\widehat{\lambda} = \|\widehat{\mathbf{\Sigma}_{noise}}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2\delta \log M}{N}}\right)^2$, where $\|\widehat{\mathbf{\Sigma}_{noise}}\|_2 = \widehat{\sigma}^2$ is the median absolute deviation (MAD) estimator of σ^2 used in standard wavelet denoising (see e.g. Antoniadis et al. [4]) and $\delta = 1.1$. Hence, the method to compute $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ is fully data-driven. Furthermore, we will show in the examples below that replacing λ by $\widehat{\lambda}$ into the penalized criterion yield a very good practical performance of the covariance estimation procedure.

As a final step, one needs to compute the estimator $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ of $\mathbf{\Sigma}$, as in (4.26). For this, we need to have an idea of the true sparsity s^*, since $\widehat{J}$ defined in (4.22) depends on s^* and also on unknown upper bounds on the level of noise in the matrix regression model (4.8). A similar problem arises in the selection of a sparse set of active variables when using regularization by a Lasso penalty in the standard linear model. Recall that in Lounici [61], a second thresholding step is used to estimate the true sparsity pattern. However, the suggested thresholding procedure in [61] also depends on a priori unknown quantities (such as the amplitude of the coefficients to estimate). To overcome this drawback in our case, we can define the final covariance group Lasso (FCGL) estimator as the matrix

$$\widehat{\mathbf{\Sigma}}_{\widehat{J}} = \mathbf{G}_{\widehat{J}} \widehat{\mathbf{\Psi}}_{\widehat{J}} \mathbf{G}_{\widehat{J}}^{\top}, \tag{4.32}$$

with $\widehat{J} = \left\{ k : \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > \varepsilon \right\}$, where ε is a positive constant. To select an appropriate value of ε we plot the sorted values $\left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2}$ of the columns of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ for $k = 1, ..., M$. Then, we use an L-curve criterion to only keep in $\widehat{J}$ the indices of the columns of $\widehat{\mathbf{\Psi}}_{\widehat{\lambda}}$ with a significant value in ℓ_2-norm. This choice for $\widehat{J}$ is sufficient for numerical purposes.

To measure the accuracy of the estimation procedure, we use the empirical averages of the Frobenius and operator norms of the estimators $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ and $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ with respect to the true covariance matrix $\mathbf{\Sigma}$ defined by $EAFN = \frac{1}{P} \sum_{p=1}^{P} \left\| \widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}^{p} - \mathbf{\Sigma} \right\|_F$ and $EAON = \frac{1}{P} \sum_{p=1}^{P} \left\| \widehat{\mathbf{\Sigma}}_{\widehat{J}}^{p} - \mathbf{\Sigma} \right\|_2$ respectively, over a number P of iterations, where $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}^{p}$ and $\widehat{\mathbf{\Sigma}}_{\widehat{J}}^{p}$ are the CGL and FCGL estimators of $\mathbf{\Sigma}$, respectively, obtained at the p-th iteration. We also compute the empirical average of the operator norm of the estimator $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ with respect to the matrix $\mathbf{\Sigma}_{J^*}$, defined by $EAON^* = \frac{1}{P} \sum_{p=1}^{P} \left\| \widehat{\mathbf{\Sigma}}_{\widehat{J}}^{p} - \mathbf{\Sigma}_{J^*} \right\|_2$.

4.4.2 Case of an orthonormal design: $n = M$

When $n = M$ the location points $t_1, ..., t_n$ are given by the equidistant grid of points $t_j = \frac{j}{M}$, $j = 1, \ldots, M$ such that the design matrix $\mathbf{G}$ is an $M \times M$ orthonormal matrix. Note that in this setting the weighs $\gamma_k = 2\|\mathbf{G}_k\|_{\ell_2} \sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^{\top})} = 2$ for all $k = 1, ..., M$.

Figures 1, 2, and 3 present the results obtained for a particular simulated sample of size $N = 25$ according to (4.31), with $n = M = 256$, $\sigma = 0.01$, $\gamma = 0.5$ and with f being any of the functions HeaviSine or Blocks. It can be observed in Figures 1(a) and 1(b) that, as expected in this high dimensional setting ($N < n$), the empirical eigenvector of $\widetilde{\mathbf{S}}$ associated to its largest empirical eigenvalue does not lead to a consistent estimator of $\mathbf{F}$.

The CGL estimator $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ is computed directly from Proposition 4.1. In Figures 2(a) and 2(b) is shown the eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ as an estimator of $\mathbf{F}$. Note that this estimator behaves poorly. The estimation considerably improves taking the FCGL estimator $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ defined in (4.32). Figures 3(a) and 3(b) illustrate the very good performance of the eigenvector associated to the largest eigenvalue of the matrix $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ as an estimator of $\mathbf{F}$.

Figures for the case $n = M$.

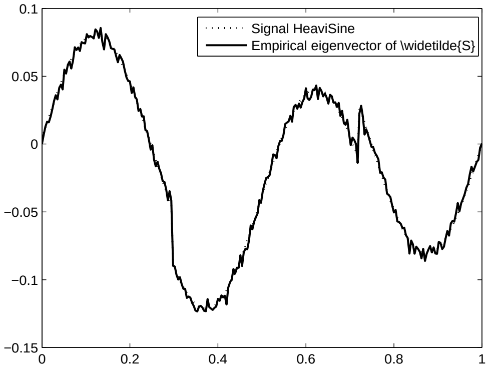


Figure 1(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathrm{S}}$.

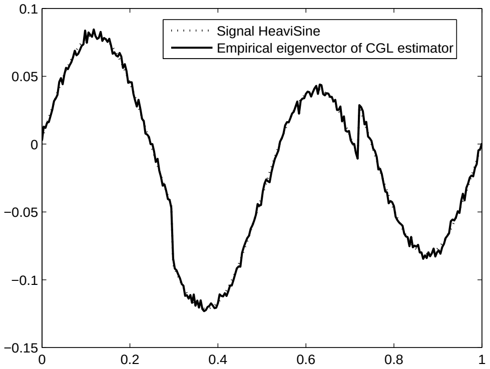


Figure 2(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$.

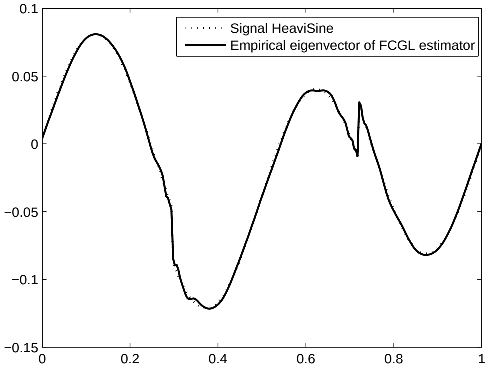


Figure 3(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$.

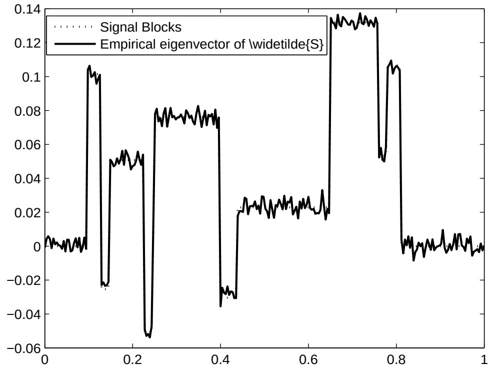


Figure 1(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathrm{S}}$.

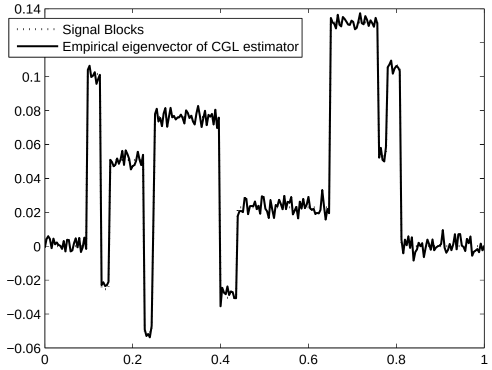


Figure 2(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$.

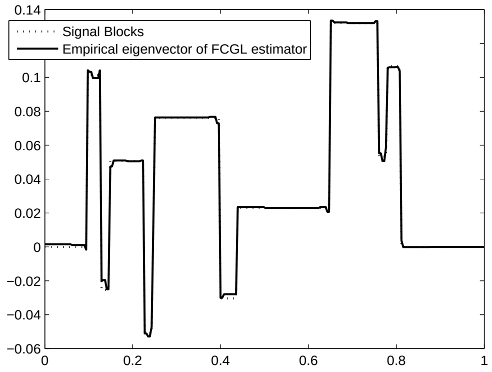


Figure 3(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$.

It is clear that the estimators $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ and $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$ are random matrices that depend on the observed sample. Tables 1(a) and 1(b) show the values of $EAFN$, $EAON$ and $EAON^*$ corresponding to $P = 100$ simulated samples of different sizes N and different values of the level of noise σ. It can be observed that for both signals the empirical averages $EAFN$, $EAON$ and $EAON^*$ behaves similarly, being the values of $EAON$ smaller than its corresponding values of $EAFN$ as expected. Observing each table separately we can remark that, for N fixed, when the level of noise σ increases then the values of $EAFN$, $EAON$ and $EAON^*$ also increase. By simple inspection of the values of $EAFN$, $EAON$ and $EAON^*$ in the same position at Tables 1(a) and 1(b) we can check that, for σ fixed, when the number of replicates N increases then the values of $EAFN$, $EAON$ and $EAON^*$ decrease in all cases. We can also observe how the difference between $EAON$ and $EAON^*$ is bigger as the level of noise increase.

Table 1(a). Values of $EAFN$, $EAON$ and $EAON^*$ corresponding to signals HeaviSine and Blocks for $M = n = 256$, $N = 25$.

Signal	σ	**0.005**	**0.01**	**0.05**	**0.1**
HeaviSine	$EAFN$	0.0634	0.0634	0.2199	0.2500
HeaviSine	$EAON$	0.0619	0.0569	0.1932	0.2500
HeaviSine	$EAON^*$	0.0619	0.0569	0.1943	0.2600
Blocks	$EAFN$	0.0553	0.0681	0.2247	0.2500
Blocks	$EAON$	0.0531	0.0541	0.2083	0.2500
Blocks	$EAON^*$	0.0531	0.0541	0.2107	0.2600

Table 1(b). Values of $EAFN$, $EAON$ and $EAON^*$ corresponding to signals HeaviSine and Blocks for $M = n = 256$, $N = 40$.

Signal	σ	**0.005**	**0.01**	**0.05**	**0.1**
HeaviSine	$EAFN$	0.0501	0.0524	0.1849	0.2499
HeaviSine	$EAON$	0.0496	0.0480	0.1354	0.2496
HeaviSine	$EAON^*$	0.0496	0.0480	0.1366	0.2596
Blocks	$EAFN$	0.0485	0.0494	0.2014	0.2500
Blocks	$EAON$	0.0483	0.0429	0.1871	0.2500
Blocks	$EAON^*$	0.0483	0.0429	0.1893	0.2600

4.4.3 Case of non equispaced design points such that $n < M$

When $n < M$ the location points are given by a subset $\{t_1, ..., t_n\} \subset \{\frac{k}{M} : k = 1, ..., M\}$ of size n, such that the design matrix $\mathbf{G}$ is an $n \times M$ matrix. For a fixed value of n, the subset $\{t_1, ..., t_n\}$ is chosen taking the first n points obtained from a random permutation of the elements of the set $\{\frac{1}{M}, \frac{2}{M}, ..., 1\}$.

Figures 4, 5, and 6 present the results obtained for a particular simulated sample of size $N = 25$ according to (4.31), with $n = 90$, $M = 128$, $\sigma = 0.02$, $\gamma = 0.5$ and with f being any of the functions HeaviSine or Blocks. It can be observed in Figures 4(a) and 4(b) that, as expected in this high dimensional setting ($N < n$), the empirical eigenvector of $\widetilde{\mathbf{S}}$ associated to its largest empirical eigenvalue are noisy versions of $\mathbf{F}$.

Figures for the case $n < M$.

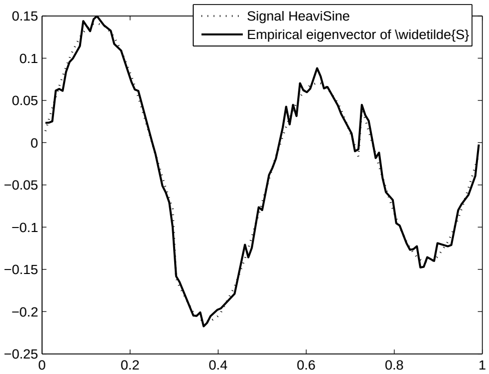


Figure 4(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathrm{S}}$.

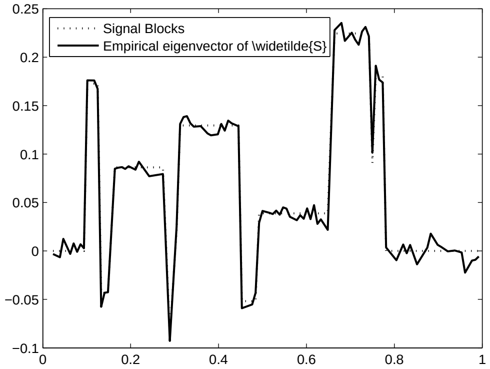


Figure 4(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widetilde{\mathrm{S}}$.

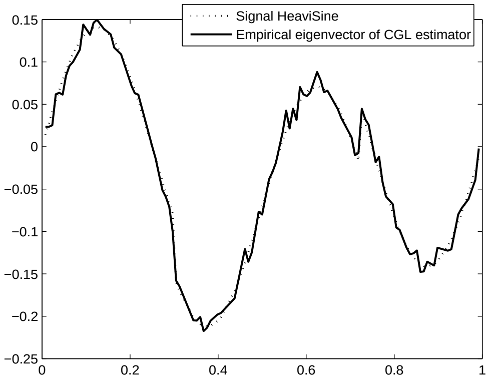


Figure 5(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$.

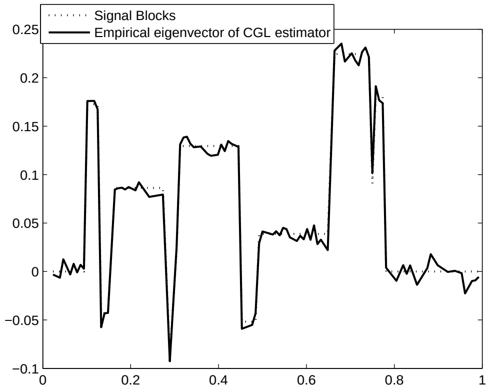


Figure 5(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$.

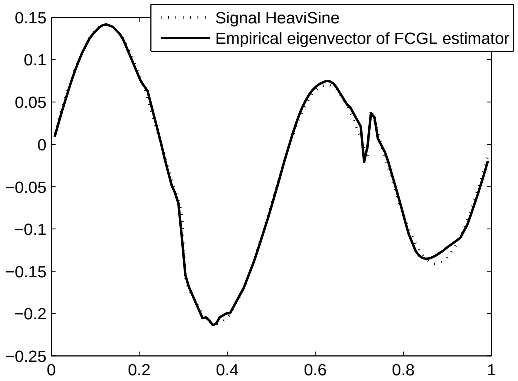


Figure 6(a). Signal HeaviSine and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$.

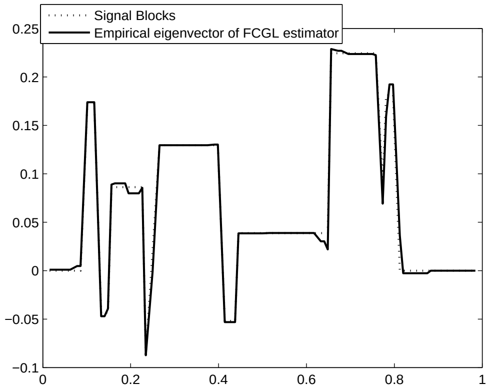


Figure 6(b). Signal Blocks and Eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{J}}$.

The CGL estimator $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ is computed by the minimization procedure (4.7) using the Matlab package *minConf* of Schmidt, Murphy, Fung and Rosales [82]. In Figures 5(a) and 5(b) is shown the eigenvector associated to the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{\widehat{\lambda}}$ as an estimator of $\mathbf{F}$. Note that while this estimator is quite noisy, the eigenvector associated to the largest eigenvalue of the matrix $\widehat{\mathbf{\Sigma}}_{\widehat{\mathcal{J}}}$ defined in (4.32) is a much better estimator of $\mathbf{F}$. This is illustrated in Figures 6(a) and 6(b).

To compare the accuracy of the estimators for different simulated samples, we compute the values of $EAFN$, $EAON$ and $EAON^*$ with fixed values of $\sigma = 0.05$, $M = 128$, $N = 40$, $P = 50$ for different values of the number of design points n. For all the values of n considered, the design points $t_1, ..., t_n$ are selected as the first n points obtained from the same random permutation of the elements of the set $\{\frac{1}{M}, \frac{2}{M} ..., 1\}$. The chosen subset $\{t_1, ..., t_n\}$ is used for all the P iterations needed in the computation of the empirical averages (fixed design over the iterations).

Figure 7 shows the values of $EAFN$, $EAON$ and $EAON^*$ obtained for each value of n for both signals HeaviSine and Blocks. It can be observed that the values of the empirical averages $EAON$ and $EAON^*$ are much smaller than its corresponding values of $EAFN$ as expected. We can remark that, when n increases, the values of $EAFN$, $EAON$ and $EAON^*$ first increase and then decrease, and the change of monotony occurs when $n > N$. Note that the case $n = M = 128$ is included in these results.

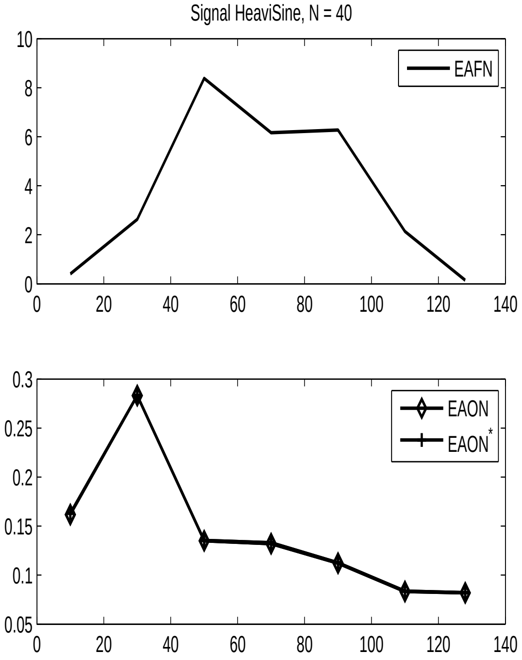


Figure 7(a). Values of $EAFN$, $EAON$ and $EAON^*$ for Signal HeaviSine as functions of n

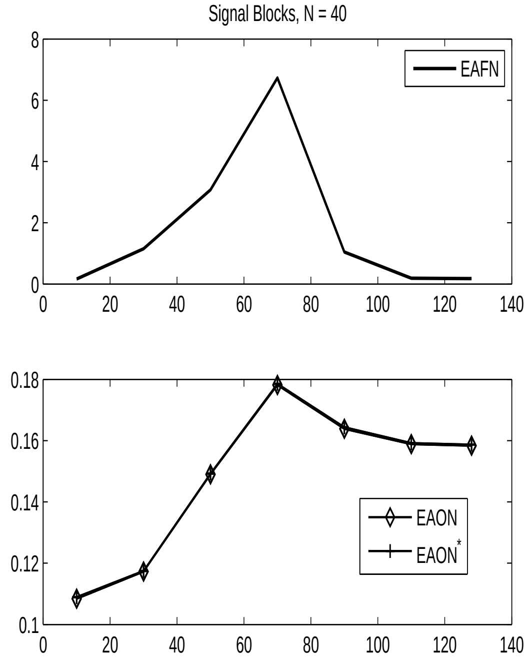


Figure 7(b). Values of $EAFN$, $EAON$ and $EAON^*$ for Signal Blocks as functions of n

4.5 Proofs

4.5.1 Proof of Proposition 4.1

Lemma 4.1. *Let $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_\lambda$ denotes the solution of (4.7). Then, for $k = 1, \ldots, M$*

$$\begin{aligned}\left[(\mathbf{G}\otimes\mathbf{G})^\top\left(vec(\widetilde{\mathbf{S}}) - (\mathbf{G}\otimes\mathbf{G})vec(\widehat{\mathbf{\Psi}})\right)\right]^k &= \lambda\gamma_k\frac{\widehat{\mathbf{\Psi}}_k}{\left\|\widehat{\mathbf{\Psi}}_k\right\|_{\ell_2}} \qquad \text{if} \quad \mathbf{\Psi}_k \neq 0 \\ \left\|\left[(\mathbf{G}\otimes\mathbf{G})^\top\left(vec(\widetilde{\mathbf{S}}) - (\mathbf{G}\otimes\mathbf{G})vec(\widehat{\mathbf{\Psi}})\right)\right]^k\right\|_{\ell_2} &\leq \lambda\gamma_k \qquad\qquad \text{if} \quad \widehat{\mathbf{\Psi}}_k = 0\end{aligned}$$

where $\widehat{\mathbf{\Psi}}_k$ denotes the k-th column of the matrix $\widehat{\mathbf{\Psi}}$ and the notation $[\beta]^k$ denotes the vector $(\beta_{k,m})_{m=1,\ldots,M}$ in $\mathbb{R}^M$ for a vector $\beta = (\beta_{k,m})_{k,m=1,\ldots,M} \in \mathbb{R}^{M^2}$.

Proof of Lemma 4.1.
For $\mathbf{\Psi} \in \mathbb{R}^{M\times M}$ define

$$L(\mathbf{\Psi}) = \left\|\widetilde{\mathbf{S}} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 = \left\|vec(\widetilde{\mathbf{S}}) - (\mathbf{G}\otimes\mathbf{G})vec(\mathbf{\Psi})\right\|_{\ell_2}^2,$$

and remark that $\widehat{\mathbf{\Psi}}$ is the solution of the convex optimization problem

$$\widehat{\mathbf{\Psi}} = \underset{\mathbf{\Psi}\in\mathcal{S}_M}{\operatorname{argmin}}\left\{L(\mathbf{\Psi}) + 2\lambda\sum_{k=1}^M\gamma_k\sqrt{\sum_{m=1}^M\Psi_{mk}^2}\right\}.$$

It follows from standard arguments in convex analysis (see e.g. Boyd and Vandenberghe [21]), that $\widehat{\mathbf{\Psi}}$ is a solution of the above minimization problem if and only if

$$-\nabla L(\widehat{\mathbf{\Psi}}) \in 2\lambda\partial\left(\sum_{k=1}^M\gamma_k\sqrt{\sum_{m=1}^M\hat{\Psi}_{mk}^2}\right),$$

where $\nabla L(\widehat{\mathbf{\Psi}})$ denotes the gradient of L at $\widehat{\mathbf{\Psi}}$ and ∂ denotes the subdifferential given by

$$\partial\left(\sum_{k=1}^M\gamma_k\sqrt{\sum_{m=1}^M\Psi_{mk}^2}\right) = \left\{\mathbf{\Theta}\in\mathbb{R}^{M\times M} : \mathbf{\Theta}_k = \gamma_k\frac{\mathbf{\Psi}_k}{\|\mathbf{\Psi}_k\|_{\ell_2}} \text{ if } \mathbf{\Psi}_k \neq 0, \|\mathbf{\Theta}_k\|_{\ell_2} \leq \gamma_k \text{ if } \mathbf{\Psi}_k = 0\right\},$$

where $\mathbf{\Theta}_k$ denotes the k-th column of $\mathbf{\Theta} \in \mathbb{R}^{M\times M}$ which completes the proof. □

Now, let $\mathbf{\Psi} \in \mathcal{S}_M$ with $M = n$ and suppose that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$. Let $\mathbf{Y} = (\mathbf{Y}_{mk})_{1\leq m,k\leq M} = \mathbf{G}^\top\widetilde{\mathbf{S}}\mathbf{G}$ and remark that $vec(\mathbf{Y}) = (\mathbf{G}\otimes\mathbf{G})^\top vec(\widetilde{\mathbf{S}})$. Then, by using Lemma 4.1 and the fact that $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_n$ implies that $(\mathbf{G}\otimes\mathbf{G})^\top(\mathbf{G}\otimes\mathbf{G}) = \mathbf{I}_{n^2}$, it follows that $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_\lambda$ satisfies for $k = 1, \ldots, M$ the following equations

$$\widehat{\mathbf{\Psi}}_k\left(1 + \frac{\lambda\gamma_k}{\sqrt{\sum_{m=1}^M\widehat{\Psi}_{mk}^2}}\right) = \mathbf{Y}_k \text{ for all } \widehat{\mathbf{\Psi}}_k \neq 0,$$

and

$$\sqrt{\sum_{m=1}^{M} \mathbf{Y}_{mk}^2} \leq \lambda \gamma_k \text{ for all } \widehat{\boldsymbol{\Psi}}_k = 0,$$

where $\widehat{\boldsymbol{\Psi}}_k = (\widehat{\Psi}_{mk})_{1 \leq m \leq M} \in \mathbb{R}^M$ and $\mathbf{Y}_k = (\mathbf{Y}_{mk})_{1 \leq m \leq M} \in \mathbb{R}^M$, which implies that the solution is given by

$$\widehat{\Psi}_{mk} = \begin{cases} 0 & \text{if } \sqrt{\sum_{m=1}^{M} \mathbf{Y}_{mk}^2} \leq \lambda \gamma_k \\ Y_{mk} \left(1 - \frac{\lambda \gamma_k}{\sqrt{\sum_{j=1}^{M} \mathbf{Y}_{jk}^2}}\right) & \text{if } \sqrt{\sum_{m=1}^{M} \mathbf{Y}_{mk}^2} > \lambda \gamma_k \end{cases}$$

which completes the proof of Proposition 4.1. □

4.5.2 Proof of Proposition 4.2

First suppose that X is Gaussian. Then, remark that for $Z = \|\mathbf{X}\|_{\ell_2}$, one has that $\|Z\|_{\psi_2} < +\infty$, which implies that $\|Z\|_{\psi_2} = \|Z^2\|_{\psi_1}^{1/2}$. Since $Z^2 = \sum_{i=1}^{n} |X(t_i)|^2$ it follows that

$$\|Z^2\|_{\psi_1} \leq \sum_{i=1}^{n} \|Z_i^2\|_{\psi_1} = \sum_{i=1}^{n} \|Z_i\|_{\psi_2}^2 = \sum_{i=1}^{n} \boldsymbol{\Sigma}_{ii} \|\boldsymbol{\Sigma}_{ii}^{-1/2} Z_i\|_{\psi_2}^2,$$

where $Z_i = X(t_i), i = 1, \ldots, n$ and $\boldsymbol{\Sigma}_{ii}$ denotes the ith diagonal element of $\boldsymbol{\Sigma}$. Then, the result follows by noticing that $\|Y\|_{\psi_2} \leq \sqrt{8/3}$ if $Y \sim N(0,1)$. The proof for the case where X is such that $\|Z\|_{\psi_2} < +\infty$ and there exists a constant C_1 such that $\|\boldsymbol{\Sigma}_{ii}^{-1/2} Z_i\|_{\psi_2} \leq C_1$ for all $i = 1, \ldots, n$ follows from the same arguments.

Now, consider the case where X is a bounded process. Since there exists a constant $R > 0$ such that for all $t \in T$, $|X(t)| \leq R$, it follows that for $Z = \|\mathbf{X}\|_{\ell_2}$ then $Z \leq \sqrt{n} R$ which implies that for any $\alpha \geq 1$, $\|Z\|_{\psi_\alpha} \leq \sqrt{n} R (\log 2)^{-1/\alpha}$, (by definition of the norm $\|Z\|_{\psi_\alpha}$) which completes the proof of Proposition 4.2. □

4.5.3 Proof of Proposition 4.4

Under the assumption that $X = X^0$, it follows that $\boldsymbol{\Sigma} = \mathbf{G} \boldsymbol{\Psi}^* \mathbf{G}^\top$ with $\boldsymbol{\Psi}^* = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top\right)$, where $\mathbf{a}$ is the random vector of $\mathbb{R}^M$ with $\mathbf{a}_m = a_m$ for $m \in J^*$ and $\mathbf{a}_m = 0$ for $m \notin J^*$. Then, define the random vector $\mathbf{a}_{J^*} \in \mathbb{R}^{J^*}$ whose coordinates are the random coefficients a_m for $m \in J^*$. Let $\boldsymbol{\Psi}_{J^*} = \mathbb{E}\left(\mathbf{a}_{J^*} \mathbf{a}_{J^*}^\top\right)$. Note that $\boldsymbol{\Sigma} = \mathbf{G}_{J^*} \boldsymbol{\Psi}_{J^*} \mathbf{G}_{J^*}^\top$ and $\mathbf{S} = \mathbf{G}_{J^*} \widehat{\boldsymbol{\Psi}}_{J^*} \mathbf{G}_{J^*}^\top$ with $\widehat{\boldsymbol{\Psi}}_{J^*} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{a}_{J^*}^i (\mathbf{a}_{J^*}^i)^\top$, where $\mathbf{a}_{J^*}^i \in \mathbb{R}^{J^*}$ denotes the random vector whose coordinates are the random coefficients a_m^i for $m \in J^*$ such that $X_i(t) = \sum_{m \in J^*} a_m^i g_m(t),\ t \in T$.

Therefore, $\widehat{\boldsymbol{\Psi}}_{J^*}$ is a sample covariance matrix of size $s^* \times s^*$ and we can control its deviation in operator norm from $\widehat{\boldsymbol{\Psi}}_{J^*}$ by using Proposition 4.3. For this we simply have to verify conditions similar to **(A1)** and **(A2)** in Assumption 4.2 for the random vector $\mathbf{a}_{J^*} = (\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1} \mathbf{G}_{J^*}^\top \mathbf{X} \in \mathbb{R}^{s^*}$. First, let $\beta \in \mathbb{R}^{s^*}$ with $\|\beta\|_{\ell_2} = 1$. Then, remark that $\mathbf{a}_{J^*}^\top \beta = \mathbf{X}^\top \widetilde{\beta}$ with $\widetilde{\beta} = \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \beta$. Since $\|\widetilde{\beta}\|_{\ell_2} \leq \left(\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\right)^{-1/2}$ and using that X satisfies Assumption 4.2 it follows that

$$\left(\mathbb{E}|\mathbf{a}_{J^*}^\top \beta|^4\right)^{1/4} \leq \rho\left(\boldsymbol{\Sigma}\right) \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right). \tag{4.33}$$

Now let $\widetilde{Z} = \|\mathbf{a}_{J^*}\|_{\ell_2} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\mathbf{X}\|_{\ell_2}$. Given our assumptions on X it follows that there exists $\alpha \geq 1$ such that

$$\|\widetilde{Z}\|_{\psi_\alpha} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|Z\|_{\psi_\alpha} < +\infty, \tag{4.34}$$

where $Z = \|\mathbf{X}\|_{\ell_2}$. Hence, using the relations (4.33) and (4.34), and Proposition 4.3 (with $\mathbf{a}_{J^*}$ instead of $\mathbf{X}$), it follows that there exists a universal constant $\delta_* > 0$ such that for all $x > 0$,

$$\mathbb{P}\left(\left\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\right\|_2 \geqslant \widetilde{\tau}_{d^*,N,s^*,1}x\right) \leqslant \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right),$$

where $\widetilde{\tau}_{d^*,N,s^*,1} = \max(\widetilde{A}^2_{d^*,N,s^*,1}, \widetilde{B}_{d^*,N,s^*,1})$, with

$$\widetilde{A}_{d^*,N,s^*,1} = \|\widetilde{Z}\|_{\psi_\alpha}\frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}},$$

$$\widetilde{B}_{d^*,N,s^*,1} = \frac{\rho^2\left(\mathbf{\Sigma}\right)\rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\sqrt{N}} + \|\mathbf{\Psi}_{J^*}\|_2^{1/2}\,\widetilde{A}_{d^*,N,s^*,1}$$

and $d^* = \min(N, s^*)$. Then, using the inequality $\|\mathbf{S}-\mathbf{\Sigma}\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\|_2$, it follows that

$$\begin{aligned}
& \mathbb{P}\left(\|\mathbf{S}-\mathbf{\Sigma}\|_2 \geq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\widetilde{\tau}_{d^*,N,s^*,1}x\right) \\
\leq\ & \mathbb{P}\left(\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\left\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\right\|_2 \geqslant \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\widetilde{\tau}_{d^*,N,s^*,1}x\right) \\
=\ & \mathbb{P}\left(\left\|\widehat{\mathbf{\Psi}}_{J^*} - \mathbf{\Psi}_{J^*}\right\|_2 \geqslant \widetilde{\tau}_{d^*,N,s^*,1}x\right) \\
\leqslant\ & \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right).
\end{aligned}$$

Hence, the result follows with

$$\begin{aligned}
\widetilde{\tau}_{d^*,N,s^*} &= \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\widetilde{\tau}_{d^*,N,s^*,1} \\
&= \max(\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\widetilde{A}^2_{d^*,N,s^*,1}, \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\widetilde{B}_{d^*,N,s^*,1}) \\
&= \max(\widetilde{A}^2_{d^*,N,s^*}, \widetilde{B}_{d^*,N,s^*}),
\end{aligned}$$

where $\widetilde{A}_{d^*,N,s^*} = \rho_{\max}^{1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\widetilde{Z}\|_{\psi_\alpha}\frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}}$ and, using the inequality

$$\|\mathbf{\Psi}_{J^*}\|_2 = \left\|\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}\mathbf{G}_{J^*}^\top \mathbf{\Sigma}\mathbf{G}_{J^*}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}\right\|_2 \leq \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}\|_2,$$

$$\widetilde{B}_{d^*,N,s^*} = \left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right)\frac{\rho^2\left(\mathbf{\Sigma}\right)}{\sqrt{N}} + \left(\frac{\rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}{\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)}\right)^{1/2}\|\mathbf{\Sigma}\|_2^{1/2}\,\widetilde{A}_{d^*,N,s^*}.$$

□

4.5.4 Proof of Theorem 4.1

Let us first prove the following lemmas.

Lemma 4.2. *Let $\mathcal{E}_1, ..., \mathcal{E}_N$ be independent copies of a second order Gaussian process $\mathcal{E}$ with zero mean. Let $\mathbf{W} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{W}_i$ with*

$$\mathbf{W}_i = \mathcal{E}_i \mathcal{E}_i^\top \in \mathbb{R}^{n \times n} \text{ and } \mathcal{E}_i = \left(\mathcal{E}_i\left(t_1\right), ..., \mathcal{E}_i\left(t_n\right)\right)^\top, \ i = 1, \ldots, N.$$

Let $\mathbf{\Sigma}_{noise} = \mathbb{E}\left(\mathbf{W}_1\right)$. For $1 \leq k \leq M$, let η_k be the k-th column of the matrix $\mathbf{G}^\top \mathbf{W} \mathbf{G}$. Then, for any $x > 0$,

$$\mathbb{P}\left(\|\eta_k\|_{\ell_2} \geq \|\mathbf{G}_k\|_{\ell_2} \sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)} \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2x}{N}}\right)^2 \right) \leq \exp(-x).$$

Proof of Lemma 4.2.

By definition one has that $\|\eta_k\|_{\ell_2}^2 = \mathbf{G}_k^\top \mathbf{W} \mathbf{G} \mathbf{G}^\top \mathbf{W} \mathbf{G}_k$ where $\mathbf{G}_k$ denotes the k-th column of $\mathbf{G}$. Hence

$$\|\eta_k\|_{\ell_2}^2 \leq \|\mathbf{G}_k\|_{\ell_2}^2 \rho_{\max}(\mathbf{G}\mathbf{G}^\top) \|\mathbf{W}\|_2^2.$$

Then, the result follows by Theorem II.13 in Davidson and Szarek [32] which can be used to prove that for any $x > 0$ then

$$\mathbb{P}\left(\|\mathbf{W}\|_2 \geq \|\mathbf{\Sigma}_{noise}\|_2 \left(1 + \sqrt{\frac{n}{N}} + \sqrt{\frac{2x}{N}}\right)^2 \right) \leq \exp(-x).$$

□

Lemma 4.3. *Let $1 \leq s \leq \min(n, M)$ and suppose that Assumption 4.1 holds for some $c_0 > 0$. Let $J \subset \{1, \ldots, M\}$ be a subset of indices of cardinality $|J| \leq s$. Let $\mathbf{\Delta} \in \mathcal{S}_M$ and suppose that*

$$\sum_{k \in J^c} \|\mathbf{\Delta}_k\|_{\ell_2} \leq c_0 \sum_{k \in J} \|\mathbf{\Delta}_k\|_{\ell_2},$$

where $\mathbf{\Delta}_k$ denotes the k-th column of $\mathbf{\Delta}$. Let

$$\kappa_{s,c_0} = \left(\rho_{\min}(s)^2 - c_0 \theta(\mathbf{G}) \rho_{\max}(\mathbf{G}^\top \mathbf{G}) s\right)^{1/2}.$$

Then,

$$\left\|\mathbf{G}\mathbf{\Delta}\mathbf{G}^\top\right\|_F^2 \geq \kappa_{s,c_0}^2 \left\|\mathbf{\Delta}_J\right\|_F^2,$$

where $\mathbf{\Delta}_J$ denotes the $M \times M$ matrix obtained by setting to zero the rows and columns of $\mathbf{\Delta}$ whose indices are not in J.

Proof of Lemma 4.3.

First let us introduce some notations. For $\mathbf{\Delta} \in \mathcal{S}_M$ and $J \subset \{1, \dots, M\}$, then $\mathbf{\Delta}_{J^c}$ denotes the $M \times M$ matrix obtained by setting to zero the rows and columns of $\mathbf{\Delta}$ whose indices are not in the complementary J^c of J. Now, remark that

$$\begin{aligned} \left\|\mathbf{G}\mathbf{\Delta}\mathbf{G}^\top\right\|_F^2 &= \left\|\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\right\|_F^2 + \left\|\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right\|_F^2 + 2\mathrm{Tr}\left(\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right) \\ &\geq \left\|\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\right\|_F^2 + 2\mathrm{Tr}\left(\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right). \end{aligned} \tag{4.35}$$

Let $\mathbf{A} = \mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top$ and $\mathbf{B} = \mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top$. Using that $\mathrm{Tr}\left(\mathbf{A}^\top\mathbf{B}\right) = vec(\mathbf{A})^\top vec(\mathbf{B})$ and the properties (A.1) and (A.55) in the Appendix, it follows that

$$\mathrm{Tr}\left(\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right) = vec(\mathbf{\Delta}_J)\left(\mathbf{G}^\top\mathbf{G} \otimes \mathbf{G}^\top\mathbf{G}\right) vec(\mathbf{\Delta}_{J^c}). \tag{4.36}$$

Let $\mathbf{C} = \mathbf{G}^\top\mathbf{G} \otimes \mathbf{G}^\top\mathbf{G}$ and note that $\mathbf{C}$ is an $M^2 \times M^2$ matrix whose elements can be written in the form of $M \times M$ block matrices given by

$$\mathbf{C}_{ij} = (\mathbf{G}^\top\mathbf{G})_{ij}\mathbf{G}^\top\mathbf{G}, \text{ for } 1 \leq i, j \leq M.$$

Now, write the $M^2 \times 1$ vectors $vec(\mathbf{\Delta}_J)$ and $vec(\mathbf{\Delta}_{J^c})$ in the form of block vectors as $vec(\mathbf{\Delta}_J) = [(\mathbf{\Delta}_J)_i^\top]_{1 \leq i \leq M}^\top$ and $vec(\mathbf{\Delta}_{J^c}) = [(\mathbf{\Delta}_{J^c})_j^\top]_{1 \leq j \leq M}^\top$, where $(\mathbf{\Delta}_J)_i \in \mathbb{R}^M$ and $(\mathbf{\Delta}_{J^c})_j \in \mathbb{R}^M$ for $1 \leq i, j \leq M$. Using (4.36) it follows that

$$\begin{aligned} \mathrm{Tr}\left(\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right) &= \sum_{1 \leq i,j \leq M} (\mathbf{\Delta}_J)_i^\top \mathbf{C}_{ij} (\mathbf{\Delta}_{J^c})_j \\ &= (\mathbf{G}^\top\mathbf{G})_{ij} \sum_{i \in J} \sum_{j \in J^c} (\mathbf{\Delta}_J)_i^\top \mathbf{G}^\top\mathbf{G} (\mathbf{\Delta}_{J^c})_j. \end{aligned}$$

Now, using that $\left|(\mathbf{G}^\top\mathbf{G})_{ij}\right| \leq \theta(\mathbf{G})$ for $i \neq j$ and that

$$\left|(\mathbf{\Delta}_J)_i^\top \mathbf{G}^\top\mathbf{G}(\mathbf{\Delta}_{J^c})_j\right| \leq \|\mathbf{G}(\mathbf{\Delta}_J)_i\|_{\ell_2} \|\mathbf{G}(\mathbf{\Delta}_{J^c})_j\|_{\ell_2} \leq \rho_{\max}(\mathbf{G}^\top\mathbf{G}) \|(\mathbf{\Delta}_J)_i\|_{\ell_2} \|(\mathbf{\Delta}_{J^c})_j\|_{\ell_2},$$

it follows that

$$\mathrm{Tr}\left(\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right) \geq -\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G}) \left(\sum_{i \in J} \|(\mathbf{\Delta}_J)_i\|_{\ell_2}\right) \left(\sum_{j \in J^c} \|(\mathbf{\Delta}_{J^c})_j\|_{\ell_2}\right).$$

Now, using the assumption that $\sum_{k \in J^c} \|\mathbf{\Delta}_k\|_{\ell_2} \leq c_0 \sum_{k \in J} \|\mathbf{\Delta}_k\|_{\ell_2}$ it follows that

$$\begin{aligned} \mathrm{Tr}\left(\mathbf{G}\mathbf{\Delta}_J\mathbf{G}^\top\mathbf{G}\mathbf{\Delta}_{J^c}\mathbf{G}^\top\right) &\geq -c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G}) \left(\sum_{i \in J} \|(\mathbf{\Delta}_J)_i\|_{\ell_2}\right)^2 \\ &\geq -c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G}) s \left\|\mathbf{\Delta}_J\right\|_F^2, \end{aligned} \tag{4.37}$$

where, for the inequality, we used the properties that for the positive reals $c_i = \|(\mathbf{\Delta}_J)_i\|_{\ell_2}$, with $i \in J$, then $\left(\sum_{i \in J} c_i\right)^2 \leq |J| \sum_{i \in J} c_i^2 \leq s \sum_{i \in J} c_i^2$ and that $\sum_{i \in J} \|(\mathbf{\Delta}_J)_i\|_{\ell_2}^2 = \|\mathbf{\Delta}_J\|_F^2$.

Using the properties (A.1) and (A.54) in the Appendix remark that

$$\begin{aligned}\left\|\mathbf{G}\boldsymbol{\Delta}_J\mathbf{G}^\top\right\|_F^2 &= \|\mathbf{G}_J\otimes\mathbf{G}_J\; vec(\tilde{\boldsymbol{\Delta}}_J)\|_{\ell_2}^2\\ &\geq \rho_{\min}\left(\mathbf{G}_J\otimes\mathbf{G}_J\right)\|vec(\tilde{\boldsymbol{\Delta}}_J)\|_{\ell_2}^2\\ &\geq \rho_{\min}(s)^2\,\|\boldsymbol{\Delta}_J\|_F^2\,, \qquad (4.38)\end{aligned}$$

where $vec(\tilde{\boldsymbol{\Delta}}_J)=[(\boldsymbol{\Delta}_J)_i^\top]_{i\in J}^\top$. Therefore, combining inequalities (4.35), (4.37) and (4.38) it follows that

$$\left\|\mathbf{G}\boldsymbol{\Delta}\mathbf{G}^\top\right\|_F^2 \geq \left(\rho_{\min}(s)^2 - c_0\theta(\mathbf{G})\rho_{\max}(\mathbf{G}^\top\mathbf{G})s\right)\|\boldsymbol{\Delta}_J\|_F^2\,,$$

which completes the proof of Lemma 4.3. □

Let us now proceed to the proof of Theorem 4.1. Part of the proof is inspired by results in Bickel, Ritov and Tsybakov [13]. Let $s\leq\min(n,M)$ and $\boldsymbol{\Psi}\in\mathcal{S}_M$ with $\mathcal{M}(\boldsymbol{\Psi})\leq s$. Let $J=\{k\;;\boldsymbol{\Psi}_k\neq 0\}$. To simplify the notations, write $\widehat{\boldsymbol{\Psi}}=\widehat{\boldsymbol{\Psi}}_\lambda$. By definition of $\widehat{\boldsymbol{\Sigma}}_\lambda=\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top$ one has that

$$\left\|\widetilde{\mathbf{S}}-\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top\right\|_F^2+2\lambda\sum_{k=1}^M\gamma_k\|\widehat{\boldsymbol{\Psi}}_k\|_{\ell_2}\leq\left\|\widetilde{\mathbf{S}}-\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^\top\right\|_F^2+2\lambda\sum_{k=1}^M\gamma_k\|\boldsymbol{\Psi}_k\|_{\ell_2}. \qquad (4.39)$$

Using the scalar product associated to the Frobenius norm $\langle\mathbf{A},\mathbf{B}\rangle_F=\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{B}\right)$ then

$$\begin{aligned}\left\|\widetilde{\mathbf{S}}-\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top\right\|_F^2 &= \left\|\mathbf{S}+\mathbf{W}-\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top\right\|_F^2\\ &= \|\mathbf{W}\|_F^2+\left\|\mathbf{S}-\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top\right\|_F^2+2\left\langle\mathbf{W},\mathbf{S}-\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top\right\rangle_F. \qquad (4.40)\end{aligned}$$

Putting (4.40) in (4.39) we get

$$\begin{aligned}\left\|\mathbf{S}-\mathbf{G}\widehat{\boldsymbol{\Psi}}\mathbf{G}^\top\right\|_F^2+2\lambda\sum_{k=1}^M\gamma_k\|\widehat{\boldsymbol{\Psi}}_k\|_{\ell_2} &\leq \left\|\mathbf{S}-\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^\top\right\|_F^2+2\left\langle\mathbf{W},\mathbf{G}\left(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\right)\mathbf{G}^\top\right\rangle_F\\ &\quad +2\lambda\sum_{k=1}^M\gamma_k\|\boldsymbol{\Psi}_k\|_{\ell_2}.\end{aligned}$$

For $k=1,\ldots,M$ define the $M\times M$ matrix $\mathbf{A}_k$ with all columns equal to zero except the k-th which is equal to $\widehat{\boldsymbol{\Psi}}_k-\boldsymbol{\Psi}_k$. Then, remark that

$$\begin{aligned}\left\langle\mathbf{W},\mathbf{G}\left(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\right)\mathbf{G}^\top\right\rangle_F &= \sum_{k=1}^M\left\langle\mathbf{W},\mathbf{G}\mathbf{A}_k\mathbf{G}^\top\right\rangle_F=\sum_{k=1}^M\left\langle\mathbf{G}^\top\mathbf{W}\mathbf{G},\mathbf{A}_k\right\rangle_F\\ &= \sum_{k=1}^M\eta_k^\top(\widehat{\boldsymbol{\Psi}}_k-\boldsymbol{\Psi}_k)\leq\sum_{k=1}^M\|\eta_k\|_{\ell_2}\|\widehat{\boldsymbol{\Psi}}_k-\boldsymbol{\Psi}_k\|_{\ell_2},\end{aligned}$$

where η_k is the k-th column of the matrix $\mathbf{G}^\top\mathbf{W}\mathbf{G}$. Define the event

$$\mathcal{A}=\bigcap_{k=1}^M\left\{2\|\eta_k\|_{\ell_2}\leq\lambda\gamma_k\right\}. \qquad (4.41)$$

Then, the choices

$$\gamma_k = 2\|\mathbf{G}_k\|_{\ell_2}\sqrt{\rho_{\max}(\mathbf{G}\mathbf{G}^\top)},\ \lambda = \|\mathbf{\Sigma}_{noise}\|_2\left(1+\sqrt{\frac{n}{N}}+\sqrt{\frac{2\delta\log M}{N}}\right)^2,$$

and Lemma 4.2 imply that the probability of the complementary event $\mathcal{A}^c$ satisfies

$$\mathbb{P}\left(\mathcal{A}^c\right) \le \sum_{k=1}^{M}\mathbb{P}\left(2\|\eta_k\|_{\ell_2} > \lambda\gamma_k\right) \le M^{1-\delta}.$$

Then, on the event $\mathcal{A}$ one has that

$$\begin{aligned}\left\|\mathbf{S}-\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 &\le \left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2+\lambda\sum_{k=1}^{M}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}\\ &\quad +2\lambda\sum_{k=1}^{M}\gamma_k\left(\|\mathbf{\Psi}_k\|_{\ell_2}-\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}\right).\end{aligned}$$

Adding the term $\lambda\sum_{k=1}^{M}\ \gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}$ to both sides of the above inequality yields on the event $\mathcal{A}$,

$$\begin{aligned}\left\|\mathbf{S}-\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2+\lambda\sum_{k=1}^{M}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2} &\le \left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2\\ &\quad +2\lambda\sum_{k=1}^{M}\gamma_k\left(\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}+\|\mathbf{\Psi}_k\|_{\ell_2}-\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}\right).\end{aligned}$$

Now, remark that for all $k\notin J$, then $\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}+\|\mathbf{\Psi}_k\|_{\ell_2}-\|\widehat{\mathbf{\Psi}}_k\|_{\ell_2}=0$, which implies that

$$\begin{aligned}&\left\|\mathbf{S}-\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2+\lambda\sum_{k=1}^{M}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}\\ \le\ &\left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2+4\lambda\sum_{k\in J}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2} &(4.42)\\ \le\ &\left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2+4\lambda\sqrt{\mathcal{M}(\mathbf{\Psi})}\sqrt{\sum_{k\in J}\gamma_k^2\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}^2}, &(4.43)\end{aligned}$$

where for the last inequality we have used again the property that for the positive reals $c_k=\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}$, $k\in J$, then $\left(\sum_{k\in J}c_k\right)^2\le\mathcal{M}(\mathbf{\Psi})\sum_{k\in J}c_k^2$.

Let $\epsilon>0$ and define the event $\mathcal{A}_1=\left\{4\lambda\sum_{k\in J}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}>\epsilon\left\|\mathbf{S}-\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2\right\}$. Note that on the event $\mathcal{A}\cap\mathcal{A}_1^c$ then the result of the theorem trivially follows from inequality (4.42). Now consider the event $\mathcal{A}\cap\mathcal{A}_1$ (all the following inequalities hold on this event). Using (4.42) one has that

$$\lambda\sum_{k=1}^{M}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}\le 4(1+1/\epsilon)\lambda\sum_{k\in J}\gamma_k\|\widehat{\mathbf{\Psi}}_k-\mathbf{\Psi}_k\|_{\ell_2}. \tag{4.44}$$

Therefore, on $\mathcal{A} \cap \mathcal{A}_1$

$$\sum_{k \notin J} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2} \leq (3 + 4/\epsilon) \sum_{k \in J} \gamma_k \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2}.$$

Let $\mathbf{\Delta}$ be the $M \times M$ symmetric matrix with columns equal to $\mathbf{\Delta}_k = \gamma_k \left(\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\right)$, for all $k = 1, \ldots, M$, and $c_0 = 3 + 4/\epsilon$. Then, the above inequality is equivalent to $\sum_{k \in J^c} \|\mathbf{\Delta}_k\|_{\ell_2} \leq c_0 \sum_{k \in J} \|\mathbf{\Delta}_k\|_{\ell_2}$ and thus Assumption 4.1 and Lemma 4.3 imply that

$$\kappa_{s,c_0}^2 \sum_{k \in J} \gamma_k^2 \|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k\|_{\ell_2}^2 \leq \left\|\mathbf{G}\mathbf{\Delta}\mathbf{G}^\top\right\|_F^2 \leq 4\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G}) \left\|\mathbf{G}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})\mathbf{G}^\top\right\|_F^2. \quad (4.45)$$

Let $\gamma_{\max}^2 = 4\mathbf{G}_{\max}^2 \rho_{\max}(\mathbf{G}^\top\mathbf{G})$. Combining the above inequality with (4.43) yields

$$\begin{aligned}\left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 &\leq \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + 4\lambda\kappa_{s,c_0}^{-1}\gamma_{\max}\sqrt{\mathcal{M}(\mathbf{\Psi})}\left\|\mathbf{G}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})\mathbf{G}^\top\right\|_F \\ &\leq \left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + 4\lambda\kappa_{s,c_0}^{-1}\gamma_{\max}\sqrt{\mathcal{M}(\mathbf{\Psi})}\left(\left\|\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top - \mathbf{S}\right\|_F + \left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{S}\right\|_F\right).\end{aligned}$$

Now, arguing as in Bickel, Ritov and Tsybakov [13], a decoupling argument using the inequality $2xy \leq bx^2 + b^{-1}y^2$ with $b > 1$, $x = 2\lambda\kappa_{s,c_0}^{-1}\gamma_{\max}\sqrt{\mathcal{M}(\mathbf{\Psi})}$ and y being either $\left\|\mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top - \mathbf{S}\right\|_F$ or $\left\|\mathbf{G}\mathbf{\Psi}\mathbf{G}^\top - \mathbf{S}\right\|_F$ yields the inequality

$$\left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 \leq \left(\frac{b+1}{b-1}\right)\left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 + \frac{8b^2\gamma_{\max}^2}{(b-1)\kappa_{s,c_0}^2}\lambda^2\mathcal{M}(\mathbf{\Psi}). \quad (4.46)$$

Then, taking $b = 1 + 2/\epsilon$ and using the inequalities $\left\|\mathbf{\Sigma} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2 \leq 2\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_F^2 + 2\left\|\mathbf{S} - \mathbf{G}\widehat{\mathbf{\Psi}}\mathbf{G}^\top\right\|_F^2$ and $\left\|\mathbf{S} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2 \leq 2\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_F^2 + 2\left\|\mathbf{\Sigma} - \mathbf{G}\mathbf{\Psi}\mathbf{G}^\top\right\|_F^2$ completes the proof of Theorem 4.1. □

4.5.5 Proof of Theorem 4.2

Part of the proof is inspired by the approach followed in Lounici [61] and Lounici, Pontil, Tsybakov and van de Geer [62]. Note first that

$$\max_{1 \leq k \leq M} \gamma_k \left\|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^*\right\|_{\ell_2} \leq \sum_{k=1}^M \gamma_k \left\|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^*\right\|_{\ell_2}.$$

Since $\mathbf{\Psi}^* \in \{\mathbf{\Psi} \in \mathcal{S}_M : M(\mathbf{\Psi}) \leq s^*\}$, we can use some results from the proof of Theorem 4.1. On the event $\mathcal{A}$ (of probability $1 - M^{1-\delta}$) defined by (4.41) and using (4.44) we get

$$\begin{aligned}\sum_{k=1}^M \gamma_k \left\|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^*\right\|_{\ell_2} &\leq 4\left(1 + \frac{1}{\epsilon}\right)\sum_{k \in J^*} \gamma_k \left\|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^*\right\|_{\ell_2} \\ &\leq 4\left(1 + \frac{1}{\epsilon}\right)\sqrt{s^*}\sqrt{\sum_{k \in J^*} \gamma_k^2 \left\|\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}_k^*\right\|_{\ell_2}^2}.\end{aligned}$$

Let $\mathbf{\Delta}^*$ be the $M \times M$ symmetric matrix with columns equal to $\mathbf{\Delta}^*_k = \gamma_k \left(\widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}^*_k \right)$, $k = 1, \ldots, M$, let $\gamma_{\max} = 2\mathbf{G}_{\max} \sqrt{\rho_{\max}(\mathbf{G}^\top \mathbf{G})}$ and $c_0 = 3 + 4/\epsilon$. Then, the above inequality and (4.45) imply that on the event $\mathcal{A}$,

$$\begin{aligned}
\sum_{k=1}^{M} \gamma_k \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}^*_k \right\|_{\ell_2} &\leq \frac{4\left(1 + \frac{1}{\epsilon}\right)\sqrt{s^*}}{\kappa_{s^*,c_0}} \left\| \mathbf{G} \mathbf{\Delta}^* \mathbf{G}^\top \right\|_F \\
&\leq \frac{4\left(1 + \frac{1}{\epsilon}\right)\sqrt{s^*}}{\kappa_{s^*,c_0}} \gamma_{\max} \left\| \mathbf{G} \left(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}^* \right) \mathbf{G}^\top \right\|_F \\
&\leq \frac{4\left(1 + \epsilon\right)\sqrt{s^*}}{\epsilon \kappa_{s^*,c_0}} \gamma_{\max} \left\| \widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma} \right\|_F \\
&\leq \frac{4\left(1 + \epsilon\right)\sqrt{s^*}}{\epsilon \kappa_{s^*,c_0}} \gamma_{\max} \sqrt{n} \sqrt{C_0\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)},
\end{aligned}$$

which implies that

$$\sum_{k=1}^{M} \frac{\|\mathbf{G}_k\|_{\ell_2}}{\sqrt{n}\mathbf{G}_{\max}} \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}^*_k \right\|_{\ell_2} \leq C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right). \tag{4.47}$$

Hence $\max\limits_{1 \leq k \leq M} \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}^*_k \right\|_{\ell_2} \leq C_1\left(\sigma, n, M, N, s^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)$ with probability at least $1 - M^{1-\delta}$, which proves the first assertion of Theorem 4.2.

Then, to prove that $\widehat{J} = J^*$ we use that $\frac{\delta_k}{\sqrt{n}} \left| \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \|\mathbf{\Psi}^*_k\|_{\ell_2} \right| \leq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k - \mathbf{\Psi}^*_k \right\|_{\ell_2}$ for all $k = 1, \ldots, M$. Then, by (4.47)

$$\left| \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}^*_k\|_{\ell_2} \right| \leq C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right),$$

which is equivalent to

$$-C_1 \leq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}^*_k\|_{\ell_2} \leq C_1, \tag{4.48}$$

where $C_1 \equiv C_1\left(n, M, N, s^*, \mathbf{S}, \mathbf{\Psi}^*, \mathbf{G}, \mathbf{\Sigma}_{noise}\right)$. If $k \in \widehat{J}$ then $\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > C_1$. Inequality $\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}^*_k\|_{\ell_2} \leq C_1$ from (4.48) imply that

$$\frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}^*_k\|_{\ell_2} \geq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - C_1 > 0,$$

where the last inequality is obtained using that $k \in \widehat{J}$. Hence $\|\mathbf{\Psi}^*_k\|_{\ell_2} > 0$ and therefore $k \in J^*$. If $k \in J^*$ then $\|\mathbf{\Psi}^*_k\|_{\ell_2} \neq 0$. Inequality $-C_1 \leq \frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} - \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}^*_k\|_{\ell_2}$ from (4.48) imply that

$$\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} + C_1 \geq \frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}^*_k\|_{\ell_2} > 2C_1,$$

where the last inequality is obtained using Assumption (4.25) on $\frac{\delta_k}{\sqrt{n}} \|\mathbf{\Psi}_k^*\|_{\ell_2}$. Hence

$$\frac{\delta_k}{\sqrt{n}} \left\| \widehat{\mathbf{\Psi}}_k \right\|_{\ell_2} > 2C_1 - C_1 = C_1$$

and therefore $k \in \widehat{J}$. This completes the proof of Theorem 4.2. □

4.5.6 Proof of Theorem 4.3

Under the assumptions of Theorem 4.3, we have shown in the proof of Theorem 4.2 that $\widehat{J} = J^*$ on the event $\mathcal{A}$ defined by (4.41). Therefore, under the assumptions of Theorem 4.3 it can be checked that on the event $\mathcal{A}$ (of probability $1 - M^{1-\delta}$)

$$\widehat{\mathbf{\Sigma}}_{\widehat{J}} = \widehat{\mathbf{\Sigma}}_{J^*} = \mathbf{G}_{J^*} \widehat{\mathbf{\Psi}}_{J^*} \mathbf{G}_{J^*}^\top,$$

with

$$\widehat{\mathbf{\Psi}}_{J^*} = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top \widetilde{\mathbf{S}} \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}.$$

Now, from the definition (4.27) of $\mathbf{\Sigma}_{J^*}$ it follows that on the event $\mathcal{A}$

$$\left\| \widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^*} \right\|_2 \leq \rho_{\max}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \left\| \widehat{\mathbf{\Psi}}_{J^*} - \Lambda_{J^*} \right\|_2, \tag{4.49}$$

where $\Lambda_{J^*} = \mathbf{\Psi}_{J^*} + (\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*})^{-1} \mathbf{G}_{J^*}^\top \mathbf{\Sigma}_{noise} \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1}$. Let $\mathbf{Y}_i = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top \widetilde{\mathbf{X}}_i$ for $i = 1, \ldots, N$ and remark that

$$\widehat{\mathbf{\Psi}}_{J^*} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_i \mathbf{Y}_i^\top \text{ with } \mathbb{E}\widehat{\mathbf{\Psi}}_{J^*} = \Lambda_{J^*}.$$

Therefore, $\widehat{\mathbf{\Psi}}_{J^*}$ is a sample covariance matrix of size $s^* \times s^*$ and we can control its deviation in operator norm from Λ_{J^*} by using Proposition 4.3. For this we simply have to verify conditions similar to **(A1)** and **(A2)** in Assumption 4.2 for the random vector $\mathbf{Y} = \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \mathbf{G}_{J^*}^\top \widetilde{\mathbf{X}} \in \mathbb{R}^{s^*}$. First, let $\beta \in \mathbb{R}^{s^*}$ with $\|\beta\|_{\ell_2} = 1$. Then, remark that $\mathbf{Y}^\top \beta = \widetilde{\mathbf{X}}^\top \widetilde{\beta}$ with $\widetilde{\beta} = \mathbf{G}_{J^*} \left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)^{-1} \beta$. Since $\|\widetilde{\beta}\|_{\ell_2} \leq \left(\rho_{\min}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\right)^{-1/2}$ it follows that

$$\left(\mathbb{E}|\mathbf{Y}^\top \beta|^4\right)^{1/4} \leq \widetilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right), \tag{4.50}$$

where $\widetilde{\rho}(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise}) = 8^{1/4}\left(\rho^4(\mathbf{\Sigma}) + \rho^4(\mathbf{\Sigma}_{noise})\right)^{1/4}$.

Now let $\widetilde{Z} = \|\mathbf{Y}\|_{\ell_2} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \|\widetilde{\mathbf{X}}\|_{\ell_2}$. Given our assumptions on the process $\widetilde{X} = X + \mathcal{E}$ it follows that there exists $\alpha \geq 1$ such that

$$\|\widetilde{Z}\|_{\psi_\alpha} \leq \rho_{\min}^{-1/2}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right)\left(\|Z\|_{\psi_\alpha} + \|W\|_{\psi_\alpha}\right) < +\infty, \tag{4.51}$$

where $Z = \|\mathbf{X}\|_{\ell_2}$, $W = \|\mathcal{E}\|_{\ell_2}$, $\mathbf{X} = \left(X(t_1), ..., X(t_n)\right)^\top$ and $\mathcal{E} = \left(\mathcal{E}(t_1), ..., \mathcal{E}(t_n)\right)^\top$. Finally, remark that

$$\|\Lambda_{J^*}\|_2 \leq \|\mathbf{\Psi}_{J^*}\|_2 + \rho_{\min}^{-1}\left(\mathbf{G}_{J^*}^\top \mathbf{G}_{J^*}\right) \|\mathbf{\Sigma}_{noise}\|_2. \tag{4.52}$$

Hence, using the relations (4.50) and (4.51), the bound (4.52) and Proposition 4.3 (with $\mathbf{Y}$ instead of $\mathbf{X}$), it follows that there exists a universal constant $\delta_* > 0$ such that for all $x > 0$

$$\mathbb{P}\left(\left\|\widehat{\mathbf{\Psi}}_{J^*} - \Lambda_{J^*}\right\|_2 \geqslant \widetilde{\tau}_{N,s^*} x\right) \leqslant \exp\left(-(\delta_*^{-1}x)^{\frac{\alpha}{2+\alpha}}\right), \tag{4.53}$$

where $\widetilde{\tau}_{N,s^*} = \max(\widetilde{A}^2_{N,s^*}, \widetilde{B}_{N,s^*})$, with

$$\widetilde{A}_{N,s^*} = \|\widetilde{Z}\|_{\psi_\alpha} \frac{\sqrt{\log d^*}(\log N)^{1/\alpha}}{\sqrt{N}},$$

$$\widetilde{B}_{N,s^*} = \frac{\widetilde{\rho}^2(\mathbf{\Sigma}, \mathbf{\Sigma}_{noise})\rho^{-1}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)}{\sqrt{N}} + \left(\|\mathbf{\Psi}_{J^*}\|_2 + \rho^{-1}_{\min}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\|\mathbf{\Sigma}_{noise}\|_2\right)^{1/2}\widetilde{A}_{N,s^*}$$

and $d^* = \min(N, s^*)$. Then, define the event

$$\mathcal{B} = \left\|\widehat{\mathbf{\Psi}}_{J^*} - \Lambda_{J^*}\right\|_2 \leqslant \widetilde{\tau}_{N,s^*}\delta_\star\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}.$$

For $x = \delta_\star\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}}$ with $\delta_\star > \delta_*$, inequality (4.53) implies that $\mathbb{P}\left(\mathcal{B}\right) \geq 1 - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$. Therefore, on the event $\mathcal{A}\cap\mathcal{B}$ (of probability at least $1 - M^{1-\delta} - M^{-\left(\frac{\delta_\star}{\delta_*}\right)^{\frac{\alpha}{2+\alpha}}}$), using inequality (4.49) and the fact that $\widehat{J} = J^*$ one obtains

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{J}} - \mathbf{\Sigma}_{J^*}\right\|_2 \leq \rho_{\max}\left(\mathbf{G}^\top_{J^*}\mathbf{G}_{J^*}\right)\widetilde{\tau}_{N,s^*}\delta_\star\left(\log(M)\right)^{\frac{2+\alpha}{\alpha}},$$

which completes the proof of Theorem 4.3. □

General Conclusions and Perspectives

In this work we have studied the problem of covariance estimation. We have proposed different nonparametric methods for the estimation of the covariance function and the covariance matrix of a stochastic process under different conditions on the process. We summarize very briefly the different results we have obtained and give some perpectives about possible future work.

Chapter 2

In the case of a Gaussian and stationary process, the covariance function estimation was done in an adaptive way, using an appropriate nonlinear wavelet thresholding procedure. The spectral density of the process was estimated by projecting the wavelet thresholding expansion of the periodogram onto a family of exponential functions to ensure that the spectral density estimator is a strictly positive function. We showed the existence with probability tending to one of a positive nonlinear estimator for the spectral density of the process which gets rise to a real covariance function estimator. The theoretical behavior of the estimator was established in terms of the rate of convergence of the Kullback-Leibler divergence over Besov classes. We have showed that the spectral density estimator reaches what we conjectured is the optimal rate of convergence without prior knowledge of the regularity of the spectral density. We also showed the good practical performance of the estimator in some numerical experiments.

In this context would be interesting to extend our results to estimators based on tapered periodogram and to prove the optimality of the rate of convergence that we have obtained.

Chapter 3

Under general moments assumptions of the process, without additional conditions such as stationarity or Gaussianity, we have proposed to use a matrix regression model for nonparametric covariance function estimation. Observing independent and indentically distributed replications of the process at fixed observation points, we have constructed an estimator of the covariance function by expanding the process onto a collection of basis functions. We have considered the covariance function of the expansion used to approximate the underlying process as an approximation of the true covariance. We pointed out that the large amount of existing approaches for function approximation provides great flexibility to our method.

The estimator proposed is obtained by the minimization of an empirical contrast function, which is exactly the sum of the squares of the residuals corresponding to a matrix linear regression model. Since the minimization is done over the linear space of symmetric matrices, it can be transformed into a classical least squares linear problem, and the computation of the estimator is therefore quite simple. For a given design matrix (or equivalently for a given number of basis functions in the expansion) we have obtained a least squares estimator of the covariance function which is non-negative definite. The choice of the number of basis functions used in the expansion is crucial since it determines the behavior of the resulting estimator. We have proposed a model selection procedure which selects the best design matrix among a collection of candidates. For this, we have applied methods and generalized results from the theory of model selection in the linear regression context.

The optimality of the procedure was proved via an oracle inequality which warrants that the best model is selected. Some numerical examples were given to illustrate the behavior of our method.

It is worth noticing that the estimator that we obtained is a real covariance function, then it can be plugged into other procedures which requires working with a covariance function. For instance the implementation of the proposed method to real data and its impact on the kriging and other applications to the prediction of a physical situation can be of great interest. Since we considered a least squares estimator of the covariance, then a possible outlook in this context is to incorporate regularization terms or constraints into the minimization of the empirical contrast function to impose desired properties for the resulting estimator, such as smoothness or sparsity conditions.

Some others extensions and open problems related to our method are:

1) Generalization of the concentration inequalities to the estimated penalty case.

2) Inclusion of noise in the observed data.

3) Consideration of data with non zero mean.

4) To study the properties of our estimators when the observation points are random in the field, since throughout this work, we have supposed that our location points are deterministic.

Chapter 4

In a high-dimensional setting, we have considered the Group-Lasso estimator of the covariance matrix of a stochastic process corrupted by an additive Gaussian noise under the assumption that the process has a sparse representation in a large dictionary of basis functions. Using a matrix regression model, we have proposed a new methodology for high-dimensional covariance matrix estimation based on empirical contrast regularization by a Group-Lasso penalty. Using such a penalty, the method selects a sparse set of basis functions in the dictionary used to approximate the process, leading to an approximation of the covariance matrix into a low dimensional space. Hence we have proposed a new method of dimension reduction for high-dimensional data, which behaves very good in practice.

The theoretical properties of such a procedure was investigated using oracle inequalities and a non-asymptotic point of view by holding fixed the number of replicates N and the location points n. The consistency of the estimator was studied in Frobenius and operator norms. Since various results existing in matrix theory show that convergence in operator norm implies convergence of the eigenvectors and eigenvalues, then consistency in operator norm is well suited for PCA applications. This was illustrated in the numerical experiments. It was also showed that in the case where the size M of the dictionary is equal to n, and the basis functions and location points are chosen such that the design matrix is orthogonal, our approach can be interpreted as a thresholding procedure of the entries of an empirical matrix.

The results obtained for the estimation of the active set of basis functions in the dictionary involve unknown quantities. A possible outlook in this context is to find a consistent estimation procedure completely derived from the data.

A useful perspective is also to impose further conditions to the Group-Lasso estimator in order to ensure that the covariance matrix estimators obtained preserve the property of non-negative definiteness.

Appendix

We give some definitions and properties of matrix operations and special matrices used throughout the thesis (see Seber [84] and references therein).

Definition A.4. *The vectorization of an $n \times k$ matrix $\mathbf{A} = (a_{ij})_{1\leq i\leq n, 1\leq j\leq k}$ is the $nk \times 1$ column vector denoted by $vec\,(\mathbf{A})$, obtained by stacking the columns of the matrix $\mathbf{A}$ on top of one another. That is, $vec(\mathbf{A}) = [a_{11}, ..., a_{n1}, a_{12}, ..., a_{n2}, ..., a_{1k}, ..., a_{nk}]^\top$.*

The following property holds.

Property A.1. *The Frobenius norm is invariant by the vec operation, meaning that*

$$\|\mathbf{A}\|_F^2 = \|vec\,(\mathbf{A})\|_{\ell_2}^2.$$

For a symmetric $n \times n$ matrix $\mathbf{A}$, the vector $vec\,(\mathbf{A})$ contains more information than necessary, since the matrix is completely determined by the lower triangular portion, that is, the $n(n+1)/2$ entries on and below the main diagonal. Hence, it can be defined the symmetrized vectorization, which corresponds to a half-vectorization, denoted by $vech(\mathbf{A})$. More precisely,

Definition A.5. *For any matrix $\mathbf{A} = (a_{ij})_{1\leq i\leq n, 1\leq j\leq n}$, define $vech(\mathbf{A})$ as the column vector of size $n(n+1)/2 \times 1$ obtained by vectorizing only the lower triangular part of $\mathbf{A}$. That is $vech(\mathbf{A}) = [a_{11}, ..., a_{n1}, a_{22}, ..., a_{n2}, ..., a_{(n-1)(n-1)}, a_{(n-1)n}, a_{nn}]^\top$.*

There exists a unique linear transformation which transforms the half-vectorization of a matrix to its vectorization and vice-versa called, respectively, the duplication matrix and the elimination matrix. For any $n \in \mathbb{N}$, the $n^2 \times n\,(n+1)\,/2$ duplication matrix is denoted by $\mathbf{D}_n$.

The Kronecker product of two matrices is defined by:

Definition A.6. *If $\mathbf{A} = (a_{ij})_{1\leq i\leq n, 1\leq j\leq k}$ is an $n \times k$ matrix and $\mathbf{B} = (b_{ij})_{1\leq i\leq p, 1\leq j\leq q}$ is a $p \times q$ matrix, then the Kronecker product of the two matrices, denoted by $\mathbf{A} \otimes \mathbf{B}$, is the $np \times kq$ block matrix*

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & . & . & . & a_{1k}\mathbf{B} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{n1}\mathbf{B} & . & . & . & a_{nk}\mathbf{B} \end{bmatrix}.$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ can be replaced by vectors in the above definition.

Property A.2. *The following properties hold:*

$$vec\,(\mathbf{ABC}) = \left(\mathbf{C}^\top \otimes \mathbf{A}\right) vec\,(\mathbf{B})\,, \tag{A.54}$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}, \tag{A.55}$$

and

$$(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top, \tag{A.56}$$

provided the above matrix products are compatible.

Recall that for a square $n \times n$ matrix $\mathbf{A}$ with real entries, $\rho_{min}(\mathbf{A})$ denotes the smallest eigenvalue of $\mathbf{A}$, and $\rho_{max}(\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}$. For $\beta \in \mathbb{R}^k$, $\|\beta\|_{\ell_2}$ denotes the usual Euclidean norm of β.

Definition A.7. *For an $n \times k$ matrix $\mathbf{A}$ with real entries, the operator norm of $\mathbf{A}$ is defined by $\|\mathbf{A}\|_2 = \sup_{\beta \in \mathbb{R}^k,\, \beta \neq 0} \frac{\|\mathbf{A}\beta\|_{\ell_2}}{\|\beta\|_{\ell_2}}$, or equivalently by $\|\mathbf{A}\|_2 = \sup_{\beta \in \mathbb{R}^k,\, \beta = 1} \|\mathbf{A}\beta\|_{\ell_2}$. In particular, if $\mathbf{A}$ is a square matrix with real entries, then $\|\mathbf{A}\|_2 = \sqrt{\rho_{max}(\mathbf{A}^\top \mathbf{A})}$.*

Definition A.8. *The spectral radius of a square $n \times n$ matrix $\mathbf{A}$ is the maximum of the absolute values of the eigenvalues of $\mathbf{A}$ and it is denoted by $\tau\,(\mathbf{A})$, that is, $\tau\,(\mathbf{A}) = \max\limits_{i=1,\ldots,n} |\rho_i(\mathbf{A})|$, where $\rho_i(\mathbf{A})$, $i = 1, \ldots, n$ are the eigenvalues of $\mathbf{A}$. Note that $\tau\,(\mathbf{A})$ need not be an eigenvalue of $\mathbf{A}$.*

Property A.3. *If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric then $\|\mathbf{A}\|_2 = \sqrt{\rho_{max}(\mathbf{A}^2)}$ and if $\mathbf{A}$ is non-negative definite then $\|\mathbf{A}\|_2 = \rho_{max}(\mathbf{A}) = \tau\,(\mathbf{A})$.*

Bibliography

[1] Adler, R. J. *An introduction to continuity, extrema, and related topics for general Gaussian processes.* Institute of Mathematical Statistics Lecture Notes–Monograph Series, 12. Institute of Mathematical Statistics, Hayward, CA, 1990.

[2] Akaike, H. Information theory and an extension of the maximum likelihood principle. *In P.N. Petrov and F. Csaki, editors, Proceedings 2nd International Symposium on Information Theory. Akademia Kiado* (1973), 267–281.

[3] Antoniadis, A., and Bigot, J. Poisson inverse problems. *Ann. Statist. 34* (2006), 2132–2158.

[4] Antoniadis, A., Bigot, J., and Sapatinas, T. Wavelet estimators in nonparametric regression: A comparative simulation study. *Journal of Statistical Software 6*, 6 (6 2001), 1–83.

[5] Bach, F. R. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res. 9* (2008), 1179–1225.

[6] Bahadur, R. R. Examples of inconsistency of maximum likelihood estimates. *Sankhya Ser. A 20* (1958), 207–210.

[7] Baraud, Y. Model selection for regression on a fixed design. *Probab. Theory Related Fields 117*, 4 (2000), 467–493.

[8] Baraud, Y. Model selection for regression on a random design. *ESAIM Probab. Statist. 6* (2002), 127–146 (electronic).

[9] Barron, A. R., and Sheu, C. H. Approximation of density functions by sequences of exponential families. *Ann. Statist. 19* (1991), 1347–1369.

[10] Bel, L. Nonparametric variogram estimator, application to air pollution data. *In GeoENV IV. Fourth European Conference on Geostatistics for Environmental Applications, X. Sanchez-Vila, J. Carrera and J.J. Gomez-Hernandez, eds, Kluwer, Dordrecht* (2002).

[11] Bickel, P. J., and Levina, E. Covariance regularization by thresholding. *Ann. Statist. 36*, 6 (2008), 2577–2604.

[12] Bickel, P. J., and Levina, E. Regularized estimation of large covariance matrices. *Ann. Statist. 36* (2008), 199–227.

[13] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist. 37*, 4 (2009), 1705–1732.

[14] Bigot, J., Biscay, R. J., Loubes, J. M., and Muñiz Alvarez, L. Nonparametric estimation of covariance functions by model selection. *Electron. J. Statist. 4* (2010), 822–855.

[15] Bigot, J., and Van Bellegem, S. Log-density deconvolution by wavelet thresholding. *Scand. J. of Statist. 36* (2009), 749–763.

[16] Birgé, L., and Massart, P. Rates of convergence for minimum contrast estimators. *Probab. Th. Relat. Fields 97* (1993), 113–150.

[17] Birgé, L., and Massart, P. *From model selection to adaptive estimation.* Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, 1997.

[18] Biscay, R. J., Díaz Frances, E., and Rodríguez, L. M. Cross–validation of covariance structures using the Frobenius matrix distance as a discrepancy function. *Journal of Statistical Computation and Simulation 58*, 3 (1997), 195–215.

[19] Biscay, R. J., Jiménez, J. C., and González, A. Smooth approximation of nonnegative definite kernels. In *Approximation and optimization in the Caribbean, II (Havana, 1993)*, vol. 8 of *Approx. Optim.* Lang, Frankfurt am Main, pp. 114–128.

[20] Bochner, S. Vorlesungen uber fouriersche integrale. *Akademische Verlagsgesellschaft* (1932).

[21] Boyd, S., and Vandenberghe, L. *Convex optimization.* Cambridge University Press, Cambridge., 2004.

[22] Briggs, W. M., and Levine, R. A. Wavelets and field forecast verification. *Monthly Weather Review 25*, 6 (1996), 1329–1341.

[23] Brillinger, D. R. *Time Series: Data Analysis and Theory.* New York: McGraw-Hill Inc., 1981.

[24] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation for gaussian regression. *Ann. Statist. 35* (2007), 1674–1697.

[25] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Sparsity oracle inequalities for the lasso. *Electron. J. Statist. 1* (2007), 169–194.

[26] Christakos, G. On the problem of permissible covariance and variogram models. *Water Resources Research. 20* (1984), 251–265.

[27] Comte, F. Adaptive estimation of the spectrum of a stationary gaussian sequence. *Bernoulli 7(2)* (2001), 267–298.

[28] Cressie, N. A. C. *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1993. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

[29] Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab. 3* (1975), 146–158.

[30] d'Aspremont, A., Bach, F., and El Ghaoui, L. Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res. 9* (2008), 1269–1294.

[31] Daubechies, I. Ten lectures on wavelets. *PA: Society for Industrial and Applied Mathematics.* (1992), 267–298.

[32] Davidson, K. R., and Szarek, S. J. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I.* North-Holland, Amsterdam, 2001, pp. 317–366.

[33] Davies, R. Asymptotic inference in stationary gaussian time-series. *Adv. Appl. Probab 5* (2001), 469–497.

[34] DeVore, R., and Lorentz, G. *Constructive Approximation.* Berlin: Springer-Verlag, 1993.

[35] Donoho, D. L., and Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81* (1994), 425–55.

[36] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. Density estimation by wavelet thresholding. *Ann. Statist. 24*, 2 (1996), 508–539.

[37] Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. Least angle regression. *Ann. Statist. 32* (2004), 407–451.

[38] El Karoui, N. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist. 36*, 6 (2008), 2717–2756.

[39] Elogne, S. N. *Non parametric estimation of smooth stationary covariance functions by interpolation methods.* PhD thesis, Université de Toulouse I, 2003.

[40] Elogne, S. N., Perrin, O., and Thomas-Agnan, C. Nonparametric estimation of smooth stationary covariance function by interpolation methods. *Stat. Infer. Stoch. Process. 11* (2008), 177–205.

[41] Fan, J., Fan, Y., and Lv, J. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147* (2008), 186–197.

[42] Fryzlewics, P., Nason, G., and von Sachs, R. A wavelet-fisz approach to spectrum estimation. *Journal of time series analysis 29(5)* (2008).

[43] Fu, W., and Knight, K. Asymptotics for lasso-type estimators. *Ann. Statist. 28* (2000), 1356–1378.

[44] Gendre, X. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat. 2* (2008), 1345–1372.

[45] Greenshtein, E., and Ritov, Y. Persistency in high dimensional linear predictorselection and the virtue of over-parametrization. *Bernoulli 10* (2004), 971–988.

[46] Guillot, G., Senoussi, R., and Monestiez, P. A positive definite estimator of the non stationary covariance of random fields. *In GeoENV2000. Third European Conference on Geostatistics for Environmental Applications, P. Monestiez, D. Allard and R. Froidevaux, eds, Kluwer, Dordrecht* (2000).

[47] Hall, P., Fisher, N., and Hoffmann, B. On the nonparametric estimation of covariance functions. *Ann. Statist. 22(4)* (1994), 2115–2134.

[48] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. *Lecture Notes in Statistics 129. Wavelets, Approximation, and Statistical Applications.* Springer, 1998.

[49] Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist. 29*, 2 (2001), 295–327.

[50] Johnstone, I. M., and Lu, A. Y. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*, 486 (2009), 682–693.

[51] Journel, A. G. Kriging in terms of projections. *J. Internat. Assoc. Mathematical Geol. 9*, 6 (1977), 563–586.

[52] Journel, A. G., and Huijbregts, C. J. *Mining Geostatistics.* Academic, London, 1978.

[53] Katul, G., and Vidakovic, B. Identification of low-dimensional energy containing/flux transporting eddy motion in the atmospheric surface layer using wavelet thresholding methods. *Journal of the Atmospheric Sciences 55* (1998), 377–389.

[54] Kollo, T., and von Rosen, D. *Advanced multivariate statistics with matrices*, vol. 579 of *Mathematics and Its Applications (New York).* Springer, Dordrecht, 2005.

[55] Koo, J. Y. Logspline deconvolution in Besov space. *Scand. J. of Statist. 26* (1999), 73–86.

[56] Lam, C., and Fan, J. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist. 37*, 6B (2009), 4254–4278.

[57] Levina, E., Rothman, A., and Zhu, J. Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat. 2*, 1 (2008), 245–263.

[58] Loubes, J. M., and Ludena, C. Adaptive complexity regularization for inverse problems. *Electron. J. Stat. 2* (2008), 661–677.

[59] Loubes, J. M., and van de Geer, S. Adaptive estimation with soft thresholding penalties. *Statist. Neerlandica 56*, 4 (2002), 454–479.

[60] Loubes, J. M., and Yan, Y. Penalized maximum likelihood estimation with l_1 penalty. *International Journal of Applied Mathematics and Statistics 14(J09)* (2009), 35–46.

[61] Lounici, K. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat. 2* (2008), 90–102.

[62] Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. Taking advantage of sparsity in multi-task learning. *COLT* (2009).

[63] Mallat, S. *A Wavelet Tour of Signal Processing, 2nd ed.* Academic Press, San Diego, 1999.

[64] Mallows, C. Some comments on Cp. *Technometrics 15* (1973), 661–675.

[65] Massart, P. *Concentration Inequalities and Model Selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003.

[66] Matheron, G. *Traite de geostatistique appliquee, Tome I: Memoires du Bureau de Recherches Geologiques et Minieres*, vol. 14. Editions Technip, Paris, 1962.

[67] Matsuo, T., Nychka, D. W., and Paul, D. Nonstationary covariance modeling for incomplete data: smoothed monte-carlo approach. *In final revision for Computational Statistics and Data Analysis* (2010).

[68] Meinshausen, N., and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Statist. 34* (2006), 1436–1462.

[69] Mendelson, S., and Pajor, A. On singular values of matrices with independent rows. *Bernoulli 12*, 5 (2006), 761–773.

[70] Nardi, Y., and Rinaldo, A. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat. 2* (2008), 605–633.

[71] Nemirovski, A. *Topics in nonparametric statistics.* In École d'été de Probabilités de Saint-Flour XXVIII-1998. Lecture Notes in Math. 1738, 2000.

[72] Neumann, M. H. Spectral density estimation via nonlinear wavelet methods for stationary non-gaussian time series. *Journal of time series analysis 17(6)* (1996), 601–633.

[73] Osborne, M., Presnell, B., and Turlach, B. On the lasso and its dual. *J. Comput. Graphic. Statist. 9* (2000), 319–337.

[74] Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica 17(4)* (2007), 1617–1642.

[75] Perrin, O., and Senoussi, R. Reducing non-stationarity random fields to stationary and isotropy using a space deformation. *Statistics and Probability Letters 48(1)* (2000), 23–32.

[76] Ramsay, J. O., and Silverman, B. W. *Functional Data Analysis.* Springer-verlag: NY, 1997.

[77] Rao, N. R., Mingo, J. A., Speicher, R., and Edelman, A. Statistical eigen-inference from large Wishart matrices. *Ann. Statist. 36*, 6 (2008), 2850–2885.

[78] Ripley, B. D. *Spatial Statistics.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.

[79] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. Sparse permutation invariant covariance estimation. *Electron. J. Stat. 2* (2008), 494–515.

[80] Sampson, P. D., and Guttorp, P. Nonparametric representation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc. 87* (1992), 108–119.

[81] Schäfer, J., and Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol. 4* (2005).

[82] Schmidt, M., Murphy, K., Fung, G., and Rosales, R. Structure learning in random fields for heart motion abnormality detection (addendum). *CVPR08* (2008).

[83] Schoenberg, I. J. Metric spaces and completely monotone functions. *Ann. of Math. 79* (1938), 811–841.

[84] Seber, G. A. F. *SA Matrix Handbook for Statisticians.* Wiley Series in Probability and Statistics. John Wiley & Sons Inc., New Jersey, 2008.

[85] Shapiro, A., and Botha, J. D. Variogram fitting with a general class of conditionally nonnegative definite functions. *Comput. Statist. Data Anal. 11* (1991), 87–96.

[86] Stein, M. L. *Interpolation of spatial data. Some theory for kriging.* Springer Series in Statistics. New York, NY: Springer. xvii, 247 p., 1999.

[87] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58* (1996), 267–288.

[88] Van de Geer, S. A. High dimensional generalized linear models and the lasso. *Ann. Statist. 36* (2008), 614–645.

[89] Vidakovic, B. *Statistical Modeling by Wavelets.* John Wiley & Sons, Inc., 1999.

[90] Yuan, M., and Lin, Y. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B 68(1)* (2006), 49–67.

[91] Zhang, C., and Huang, J. Model-selection consistency of the lasso in highdimensional regression. *Ann. Statist. 36* (2008), 1567–1594.

[92] Zhao, P., and Yu, B. On model selection consistency of lasso. *J. Mach. Learn. Res. 7* (2006), 2541–2563.

[93] Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Statist. 15*, 2 (2006), 265–286.