Cohérence spectrale et test de corrélation en grande dimension

Alexis Rosuel[1], Pascal Vallet[2], Philippe Loubaton[1],

[1]Laboratoire IGM (CNRS, Université Gustave Eiffel)
5 Boulevard Descartes, 77454, Marne-la-Vallée, France

[2]Laboratoire IMS (CNRS, Université de Bordeaux, Bordeaux INP)
351, cours de la Libération, 33405 Talence, France

{alexis.rosuel,philippe.loubaton}@univ-eiffel.fr, pascal.vallet@bordeaux-inp.fr

Résumé – Dans cet article, nous étudions un test de corrélation pour les composantes d'une série temporelle gaussienne complexe multivariée de dimension M. Une statistique de test basée sur un estimateur empirique du maximum de la cohérence spectrale est proposée, et son risque de 1ère espèce est étudié dans un régime asymptotique des grandes dimensions où $M \to \infty$. Nous donnons également quelques résultats numériques illustrant les performances de la statistique proposée dans divers scénarios.

Abstract – In this paper, we study a correlation test for the entries of a M-variate complex Gaussian time series. A test statistic based on a sample estimate of the maximum of the spectral coherence is proposed, and the type I error is studied in a high-dimensional asymptotic regime in which $M \to \infty$. Numerical simulations are provided to evaluate the performance of the proposed statistic under different scenarios.

1 Introduction

Détecter la présence de corrélation entre plusieurs séries temporelles est un problème fondamental en traitement du signal, dont les applications courantes concernent par exemple le traitement multi-capteurs, le radar ou encore les télécommunications.

Dans ce travail, nous considérons le cas où l'on observe M séries temporelles $(y_{1,n})_{n\in\mathbb{Z}}, \dots, (y_{M,n})_{n\in\mathbb{Z}}$, supposées conjointement gaussiennes complexes circulaires et stationnaires. En notant $r_{i,j}(h) = \mathbb{E}[y_{i,n+h}\overline{y}_{j,n}]$, on s'intéresse ici à déterminer la présence de corrélations entre les M séries, qui se formalise comme le test binaire dont l'hypothèse nulle est donnée par :

$$\mathcal{H}_0 : \; r_{i,j}(h) = 0 \quad \forall\, h \in \mathbb{Z}, 1 \leq i < j \leq M. \qquad (1)$$

Le test (1) a fait l'objet de nombreux travaux dans le cadre des approches dites « temporelles », pour lesquelles des statistiques de test basées sur des estimées empiriques de la fonction de covariance $r_{i,j}$ ont été proposées, voir par exemple [3, 4, 7].

De manière équivalente, le test (1) peut être reformulé dans le domaine spectral ; définissons à ce titre pour tout $\nu \in [0, 1]$,

$$s_{i,j}(\nu) = \sum_{h\in\mathbb{Z}} r_{i,j}(h)\mathrm{e}^{-2\mathrm{i}\pi\nu h},$$

la densité spectrale et

$$c_{i,j}(\nu) = \frac{s_{i,j}(\nu)}{\sqrt{s_{i,i}(\nu)s_{j,j}(\nu)}},$$

la cohérence spectrale associées aux séries $(y_{i,n})_{n\in\mathbb{Z}}$ et $(y_{j,n})_{n\in\mathbb{Z}}$. On aboutit alors au test équivalent d'hypothèse nulle :

$$\mathcal{H}_0 : c_{i,j}(\nu) = 0 \quad \forall\nu \in [0, 1], 1 \leq i < j \leq M. \qquad (2)$$

Les approches dites « fréquentielles » pour résoudre (2) utilisent quant à elles des statistiques de test basées sur des estimées empiriques de $s_{i,j}$ ou $c_{i,j}$, voir par exemple [8, 2, 9]. Le cadre d'étude de ces travaux se place dans le régime asymptotique dit des « petites dimensions » où M est supposé fixe et où le nombre d'observations N de chaque série tend vers l'infini, et les méthodes développées sont pertinentes dans un contexte pratique où $M \ll N$.

De nombreuses applications actuelles mettant en jeu des données de grande dimension, il apparaît nécessaire de considérer un régime asymptotique des « grandes dimensions » où M tend également vers l'infini. Dans ce contexte, les travaux autour du test (2) sont encore peu nombreux. Dans [5], une statistique de test basée sur un périodogramme lissé est proposée, et étudiée dans le régime des grandes dimensions. Notons ξ_i la transformée de Fourier normalisée de $(y_{i,n})_{n=1,\dots,N}$ donnée par

$$\xi_i(\nu) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} y_{i,n}\mathrm{e}^{-\mathrm{i}2\pi\nu(n-1)},$$

ainsi que $\hat{s}_{i,j}$ l'estimateur lissé de $s_{i,j}$ donné par

$$\hat{s}_{i,j}(\nu) = \frac{1}{B+1} \sum_{b=-\frac{B}{2}}^{\frac{B}{2}} \xi_i\left(\nu + \frac{b}{N}\right)\overline{\xi_j\left(\nu + \frac{b}{N}\right)},$$

où B est un entier pair représentant la largeur de la fenêtre de lissage fréquentiel. Notons également

$$\hat{c}_{i,j}(\nu) = \frac{\hat{s}_{i,j}(\nu)}{\sqrt{\hat{s}_{i,i}(\nu)\hat{s}_{j,j}(\nu)}},$$

l'estimateur associé de $c_{i,j}(\nu)$ et $\hat{\mathbf{C}}(\nu) = (\hat{c}_{i,j}(\nu))_{i,j=1,\dots,M}$ l'estimateur de la matrice de cohérence spectrale $\mathbf{C}(\nu) = (c_{i,j}(\nu))_{i,j=1,\dots,M}$. Les travaux de [5] montrent alors que dans le régime asymptotique où $M, B, N \to \infty$ tel que $\frac{M}{B} \to c > 0$ et $\frac{B}{N} \to 0$, sous l'hypothèse nulle $\mathcal{H}_0$, la distribution empirique des valeurs propres $\left(\lambda_1(\hat{\mathbf{C}}(\nu))\right)_{m=1,\dots,M}$ de $\hat{\mathbf{C}}(\nu)$ converge vers la loi de Marcenko-Pastur μ, i.e.

$$\frac{1}{M}\sum_{k=1}^{M} \varphi\left(\lambda_k\left(\hat{\mathbf{C}}(\nu)\right)\right) \xrightarrow{\text{p.s.}} \int_{\mathbb{R}} \varphi(x)\mathrm{d}\mu(x),$$

pour toute fonction continue bornée φ. En exploitant le fait que sous certaines hypothèses alternatives, le comportement asymptotique de la distribution empirique des valeurs propres de $\hat{\mathbf{C}}(\nu)$ dévie de la loi de Marcenko-Pastur, une statistique de test basée sur les valeurs propres de $\hat{\mathbf{C}}(\nu)$ et la loi μ est proposée.

Dans cet article, nous explorons une approche alternative, non plus basée sur les valeurs propres de $\hat{\mathbf{C}}(\nu)$, mais sur l'étude du maximum des entrées non diagonales de $\hat{\mathbf{C}}(\nu)$, i.e. $\max\left\{|\hat{c}_{i,j}(\nu)|^2 : 1 \le i < j \le M, \nu \in \mathcal{V}\right\}$, où $\mathcal{V}$ est un sous-ensemble fini de $[0,1]$ que nous détaillons en section suivante. En particulier, après un recentrage et une renormalisation convenables, nous montrons [1] que cette statistique converge en loi, sous l'hypothèse nulle $\mathcal{H}_0$ et dans le régime où $M, B, N \to \infty$ décrit précédemment, vers la distribution de Gumbel, et exploitons ce résultat pour construire une statistique de test dont le risque de 1ère espèce asymptotique est contrôlé. Notons qu'une étude similaire a été conduite dans [9], dans le régime des petites dimensions ($M = 2$), en utilisant une estimée différente de $\hat{c}_{i,j}(\nu)$.

2 Hypothèses et résultats

Commençons par présenter formellement les hypothèses utilisées dans notre étude. On considère une suite de séries temporelles $(y_{m,n})_{n\in\mathbb{Z}}$, $m \ge 1$, mutuellement indépendantes et de loi gaussienne complexe circulaire. Notons r_m (au lieu de $r_{m,m}$) la fonction de covariance de $(y_{m,n})_{n\in\mathbb{Z}}$, pour laquelle on formule l'hypothèse de « mémoire courte » suivante.

Hypothèse 1. *Les fonctions de covariance $(r_m)_{m\ge1}$ vérifient*

$$\sup_{m\ge1}\sum_{h\in\mathbb{Z}}(1+|h|)\,|r_m(h)| < \infty.$$

1. Les détails de la preuve sont disponibles dans [6].

En notant s_m (au lieu de $s_{m,m}$) la densité spectrale de $(y_{m,n})_{n\in\mathbb{Z}}$, l'hypothèse 1 assure en particulier que s_m est uniformément bornée au sens où

$$\sup_{m\ge1}\max_{\nu\in[0,1]} s_m(\nu) < \infty.$$

Afin que la cohérence spectrale $c_{i,j}$ soit bien définie, il est également nécessaire de garantir que les densités spectrales $(s_m)_{m\ge1}$ ne s'annulent pas.

Hypothèse 2. *Les densités spectrales $(s_m)_{m\ge1}$ vérifient*

$$\inf_{m\ge1}\min_{[0,1]} s_m(\nu) > 0.$$

Notons que les hypothèses 1 et 2 sont standards et sont notamment vérifiées par les processus ARMA dont le filtre générateur n'admet pas de pôle ou zéro sur le cercle unité. L'hypothèse suivante donne le régime des grandes dimensions adopté dans ce travail, en précisant les rythmes de croissance des paramètres M, N, B.

Hypothèse 3. *$B = B(N)$, $M = M(N)$ sont fonctions de N tel qu'il existe $C_1, C_2 > 0$, $\rho \in]0,1[$ et $c > 0$ tel que*

$$C_1 N^\rho \le B, M \le C_2 N^\rho$$

et $\frac{M}{B} \to c > 0$ quand $N \to \infty$.

L'hypothèse 3 appelle à quelques justifications. Dans un régime asymptotique où $M, N \to \infty$ de telle sorte que $\frac{M}{N} \to 0$, on peut montrer que $\hat{\mathbf{C}}(\nu)$ est un estimateur consistant de $\mathbf{C}(\nu)$ (pour la norme spectrale) sous réserve que $\frac{B}{M} \to 0$ et moyennant quelques hypothèses supplémentaires. En pratique, ce régime asymptotique traduit des situations où l'on peut choisir B tel que $M \ll B \ll N$. Néanmoins, lorsque la dimension M est potentiellement grande et que la taille N de l'échantillon est limitée, un tel choix de B devient complexe, et il apparaît alors plus raisonnable de choisir que B du même ordre de grandeur que M, ce que modélise l'hypothèse 3.

Muni des hypothèses précédentes, nous sommes à présent en mesure de présenter le résultat principal de cet article. Définissons $\mathcal{V} = \{k\frac{B+1}{N} : k \in \mathbb{N}, k \le \frac{N}{B+1}\}$ comme le sous-ensemble des fréquences de Fourier $\{\frac{k}{N} : k \in \mathbb{N}, k \le N-1\}$ espacées de $\frac{B+1}{N}$.

Théorème 1. *Sous les hypothèses 1, 2 et 3, on a*

$$\mathbb{P}\left(\max_{\substack{1\le i<j\le M\\ \nu\in\mathcal{V}}} |\hat{c}_{i,j}(\nu)|^2 \le \frac{t+\gamma}{B+1}\right) \xrightarrow[N\to\infty]{} \mathrm{e}^{-\mathrm{e}^{-t}}$$

pour tout $t \in \mathbb{R}$, où $\gamma = \log\left(\frac{N}{B+1}\right) + \log\left(\frac{M(M-1)}{2}\right)$.

Le théorème 1 montre donc que le maximum de la cohérence spectrale $\max\left\{|\hat{c}_{i,j}(\nu)|^2 : 1 \le i < j \le M, \nu \in \mathcal{V}\right\}$, après renormalisation par $B+1$ et recentrage par le terme γ, converge en loi vers la distribution de Gumbel. Le théorème 1 implique directement via l'hypothèse 3 que

$$\max_{\substack{1\le i<j\le M\\ \nu\in\mathcal{V}}} |\hat{c}_{i,j}(\nu)|^2 = \mathcal{O}_{\mathbb{P}}\left(\frac{\log(N)}{B}\right).$$

Cette erreur d'estimation est à comparer directement au résultat classique $|\hat{c}_{i,j}(\nu)|^2 = \mathcal{O}_{\mathbb{P}}(B^{-1})$ pour i, j, ν fixés (voir [1]), et on remarque que le maximum sur i, j, ν induit une perte logarithmique en $\log(N)$. A ce titre, notons que les 2 termes logarithmiques contenus dans γ sont directement liés aux ensembles de maximisation $\mathcal{V}$ et $\{(i,j) : 1 \leq i < j \leq M\}$, dont l'ordre de grandeur de leurs cardinaux sont $\mathcal{O}(\frac{N}{B})$ et $\mathcal{O}(M^2)$ respectivement.

De plus, le théorème 1 peut être exploité pour construire une statistique de test dont le risque de 1ère espèce est contrôlé dans le régime asymptotique des grandes dimensions. En effet, considérons le seuil

$$\eta(\alpha) = \frac{\gamma - \log\log\left(\frac{1}{1-\alpha}\right)}{B+1},$$

obtenu par inversion de la f.d.r. de la loi de Gumbel, ainsi que la statistique

$$T = \mathbb{1}_{]\eta(\alpha),\infty[}\left(\max_{\substack{1\leq i<j\leq M \\ \nu\in\mathcal{V}}} |\hat{c}_{i,j}(\nu)|^2\right). \tag{3}$$

Le théorème 1 montre alors que sous l'hypothèse $\mathcal{H}_0$,

$$\mathbb{P}(T=1) \xrightarrow[N\to\infty]{} \alpha,$$

i.e. la statistique de test T est de niveau asymptotique α.

3 Illustrations numériques

Nous présentons ici quelques simulations autour du théorème 1 et des performances de la statistique de test T définie en (3), en considérant divers scénarios pour les hypothèses nulle $\mathcal{H}_0$ et alternative $\mathcal{H}_1$. On considère pour la suite que les séries temporelles suivent un modèle AR(1) multivarié, i.e. en notant $\mathbf{y}_n = (y_{1,n}, \ldots, y_{M,n})^T$,

$$\mathbf{y}_n = \mathbf{A}\mathbf{y}_{n-1} + \boldsymbol{\epsilon}_n \tag{4}$$

où $(\boldsymbol{\epsilon}_n)_{n\in\mathbb{Z}}$ est un bruit blanc $\mathcal{N}_{\mathbb{C}^M}(\mathbf{0}, \mathbf{I})$ et où $\mathbf{A}$ sera ajustée suivant des scénarios précisés ci-dessous.

Pour illustrer le théorème 1, nous considérons dans un premier temps $\mathbf{A} = \mathrm{diag}(\theta_1, \ldots, \theta_M)$ avec $\theta_1, \ldots, \theta_M$ des variables distribuées uniformément sur le disque centré de rayon 0.9. La FIGURE 1 représente le tracé de la fonction de répartition (f.d.r.) empirique de $\max_{i,j,\nu}(B+1)|\hat{c}_{i,j}(\nu)|^2 - \gamma$ pour $N = 20000, M = 500, B = 1000$, évaluée sur 10000 tirages du modèle (4), et montre une bonne adéquation avec la loi de Gumbel. La TABLE 1 représente quant à elle, toujours pour la même matrice $\mathbf{A}$, la quantité $\mathbb{P}(T=1)$ (évaluée sur 30000 tirages) pour un niveau α fixé à 5% et un jeu de valeurs pour (M, N, B) tel que $\rho = 0.7$ et $c = 0.5$ dans l'hypothèse 3. On constate que les valeurs de $\mathbb{P}(T=1)$ sont effectivement proches de α à partir de $M = 130$.

Illustrons à présent les performances du test en considérant les scénarios suivants sous $\mathcal{H}_0$ et sous deux hypothèses alternatives dites « locale » et « globale » définies comme suit. En notant $\mathbf{e}_1, \ldots, \mathbf{e}_M$ la base canonique de $\mathbb{R}^M$,

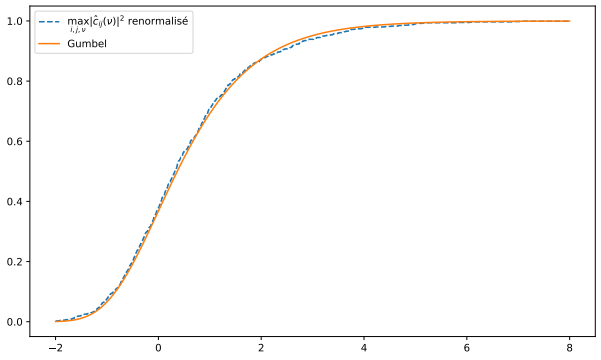

FIGURE 1 – Comparaison entre la f.d.r. empirique du maximum de la cohérence spectrale et la f.d.r. de la loi de Gumbel.

N	B	M	$\mathbb{P}(T=1)$
42	20	10	0.012
316	100	50	0.035
659	180	90	0.037
1044	260	130	0.045
1459	340	170	0.046
1901	420	210	0.048
5623	1000	500	0.048
13374	2000	1000	0.049

TABLE 1 – Risque de 1ère espèce pour $\alpha = 0.05$ en fonction de M, B, N

- sous $\mathcal{H}_0$, on considère $\mathbf{A} = \theta\,\mathbf{I}$, de telle sorte que les séries $(y_{m,n})_{n\in\mathbb{Z}}$ sont indépendantes ;
- sous $\mathcal{H}_{1,\mathrm{loc}}$, on considère $\mathbf{A} = \theta\,\mathbf{I} + \beta\,\mathbf{e}_2\mathbf{e}_1^T$, de telle sorte que $(y_{1,n})_{n\in\mathbb{Z}}$ et $(y_{2,n})_{n\in\mathbb{Z}}$ constituent l'unique couple corrélé parmi les M séries ;
- sous $\mathcal{H}_{1,\mathrm{glob}}$, on considère $\mathbf{A} = \theta\,\mathbf{I} + \beta\sum_{i-j=1}\mathbf{e}_i\mathbf{e}_j^T$, de sorte que les M séries sont corrélées deux à deux.

On choisira $\theta = 0.5$ sous les trois hypothèses et $\beta = 0.1$ sous $\mathcal{H}_{1,\mathrm{loc}}$. Afin que les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_{1,\mathrm{glob}}$ ne soient pas trivialement séparables quand $M \to \infty$, on choisit sous $\mathcal{H}_{1,\mathrm{glob}}$ le paramètre β fonction de M de sorte que le ratio r ci-dessous reste constant et égal à 0.1 :

$$r = \frac{\sum_{i\neq j}\int_0^1 |s_{i,j}(\nu)|^2 \mathrm{d}\nu}{\sum_{i,j}\int_0^1 |s_{i,j}(\nu)|^2 \mathrm{d}\nu}.$$

Nous prenons comme point de comparaison la statistique de test développée dans [5] donnée par

$$S = \mathbb{1}_{]\kappa,+\infty[}(\Delta),$$

où

$$\Delta = \max_{\nu\in[0,1]}\left|\frac{1}{M}\sum_{k=1}^{M}\varphi\left(\lambda_k\left(\hat{\mathbf{C}}(\nu)\right)\right) - \int_{\mathbb{R}}\varphi(x)\mathrm{d}\mu(x)\right|,$$

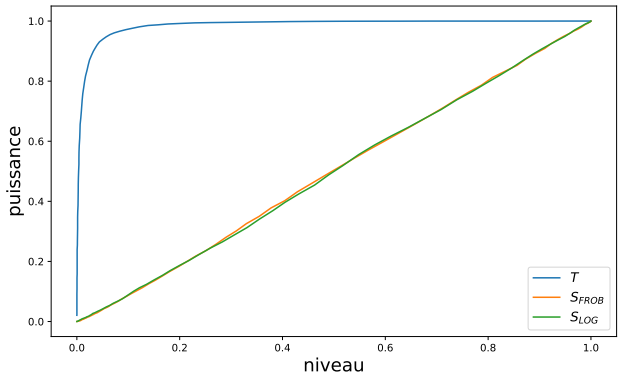

FIGURE 2 – Courbes ROC de $T, S_{\log}, S_{\text{frob}}$ pour $\mathcal{H}_{1,\text{loc}}$

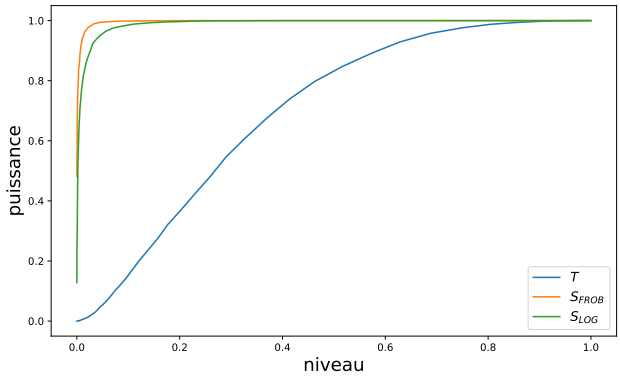

FIGURE 3 – Courbes ROC de $T, S_{\log}, S_{\text{frob}}$ pour $\mathcal{H}_{1,\text{glob}}$

et où μ est la loi de Marcenko-Pastur donnée par

$$\mathrm{d}\mu(x) = \left(1-\frac{1}{c}\right)^{+}\delta_0(\mathrm{d}x) + \frac{\sqrt{(x^+-x)(x-x^-)}}{2\pi cx}\mathbb{1}_{[x^-,x^+]}(x)\mathrm{d}x,$$

avec $x^{\pm} = (1 \pm \sqrt{c})^2$. On considèrera les fonctions $\varphi(x) = \log(x)$ et $\varphi(x) = (x-1)^2$, et les statistiques de test associées seront notées $S_{\log}$ et S_{frob}. Comme la loi asymptotique de S sous $\mathcal{H}_0$ n'est pas connue, on fixera le seuil κ de telle sorte à contrôler le risque de 1ère espèce empirique évalué par simulation Monte-Carlo (10000 tirages). Par mesure d'équité, on procédera de même pour fixer le seuil η de la statistique T définie en (3).

Les FIGURES 2 et 3 fournissent les courbes ROC empiriques des statistiques $T, S_{\log}, S_{\text{frob}}$ où $M = 290, N = 2846, B = 580$. Comme attendu, on observe que sous $\mathcal{H}_{1,\text{loc}}$, la statistique T présente les meilleures performances, puisque l'opération de maximum utilisée dans T est spécialement indiquée pour détecter les entrées non nulles de la cohérence spectrale $\mathbf{C}(\nu)$ dans un contexte parcimonieux. A contrario, sous $\mathcal{H}_{1,\text{glob}}$ où les corrélations sont réparties sur tous les couples de séries, les statistiques basées sur les valeurs propres de $\hat{\mathbf{C}}(\nu)$ sont plus adaptées. Nous présentons également en TABLES 2 et 3 la puissance des tests sous $\mathcal{H}_{1,\text{loc}}$ et $\mathcal{H}_{1,\text{glob}}$ (évaluées sur 10000 tirages)

N	M	B	S_{frob}	$S_{\log}$	T
42	10	20	0.049	0.049	0.061
316	50	100	0.038	0.044	0.352
659	90	180	0.038	0.041	0.881
1044	130	260	0.034	0.038	0.999
1459	170	340	0.034	0.038	1.000
1901	210	420	0.035	0.039	1.000
2364	250	500	0.031	0.039	1.000
2846	290	580	0.032	0.036	1.000

TABLE 2 – Puissance des statistiques $T, S_{\log}, S_{\text{frob}}$ sous $\mathcal{H}_{1,\text{loc}}$ à un niveau 5%

N	M	B	S_{frob}	$S_{\log}$	T
42	10	20	0.050	0.049	0.052
316	50	100	0.036	0.042	0.067
659	90	180	0.067	0.065	0.086
1044	130	260	0.142	0.122	0.133
1459	170	340	0.339	0.255	0.214
1901	210	420	0.601	0.462	0.328
2364	250	500	0.836	0.682	0.503
2846	290	580	0.960	0.852	0.672

TABLE 3 – Puissance des statistiques $T, S_{\log}, S_{\text{frob}}$ sous $\mathcal{H}_{1,\text{glob}}$ à un niveau 5%

pour différentes valeurs de M, N, B et un niveau fixé à 5 %. Des conclusions similaires peuvent être tirées quant aux performances des statistiques sur les deux scénarios.

Références

[1] David R Brillinger. *Time series : data analysis and theory*. SIAM, 2001.

[2] M. Eichler. Testing nonparametric and semiparametric hypotheses in vector stationary processes. *J. Multivar. Anal.*, 99(5) :968–1009, 2008.

[3] L. Haugh. Checking the independence of two covariance-stationary time series : a univariate residual cross-correlation approach. *J. Am. Stat. Assoc.*, 71(354) :378–385, 1976.

[4] Y. Hong. Testing for independence between two covariance stationary time series. *Biometrika*, 83(3) :615–625, 1996.

[5] P. Loubaton and A. Rosuel. Properties of linear spectral statistics of frequency-smoothed estimated spectral coherence matrix of high-dimensional gaussian time series. *Electron. J. Stat.*, 15(2) :5380–5454, 2021.

[6] P. Loubaton, A. Rosuel, and P. Vallet. On the asymptotic distribution of the maximum sample spectral coherence of gaussian time series in the high dimensional regime. *submitted*, 2021. arXiv :2107.02891.

[7] D. Ramírez, J. Vía, I. Santamaría, and L. Scharf. Detection of spatially correlated gaussian time series. *IEEE Trans. Signal Process.*, 58(10) :5006–5015, 2010.

[8] G. Wahba. Some tests of independence for stationary multivariate time series. *J. R. Stat. Soc. B*, 33(1) :153–166, 1971.

[9] W. Wu and P. Zaffaroni. Asymptotic theory for spectral density estimates of general multivariate time series. *Econom. Theory*, 34(1) :1–22, 2018.