QUATORZIEME COLLOQUE GRETSI - JUAN-LES-PINS - DU 13 AU 16 SEPTEMBRE 1993

Un Test de Normalité pour les Séries Temporelles Scalaires

K. Choukri et E. Moulines

ENST, Télécom Paris, Département Signal, 46 rue Barrault, 75634 Paris Cedex 13, France.

E-mail: choukri@sig.enst.fr, Tel: 33 1 45817596, Fax: 33 1 45887935.

RÉSUMÉ

Nous présentons dans cette contribution, un formalisme général concernant les procédures temporelles pour tester la normalité d'un processus stochastique stationnaire. Les formes explicites de la distribution asymptotique de ces tests statistiques sous l'hypothèse de Gaussianité, sont aussi exhibées. Deux procédures de test (les tests SK et ECF) sont comparées dans le cadre d'un exemple d'application typique: La détection d'un processus non-Gaussien additif, noyé dans un bruit stationnaire Gaussien de covariance inconnue.

ABSTRACT

In this contribution, a general formalism for time-domain procedures for testing that a stationary time-series is Gaussian, is presented. Closed-form expressions of the asymptotic distribution of the test statistics under the null hypothesis of Gaussianity are derived. Two procedures (SK and ECF-based tests) are then compared and assesed in a typical example of application: the detection of additive non-Gaussian outliers in a stationary Gaussian noise with unknown covariance.

1 Introduction

Un grand nombre de procédures permettant de tester la Gaussianité d'échantillons indépendants (monovariables ou multivariables) ont été développées ces dernières années (voir les travaux de Csörgo [1]). Par contre, peu de tentatives ont été entreprises pour tester si un processus stationnaire (scalaire ou vectoriel) est Gaussien, ceci malgré l'importance pratique de ce problème (cf. [4]). Nous pouvons citer comme exemples typiques la sélection de méthodes d'inférences appropriées, la détection de dépendance non-linéaire dans une série temporelle [6] et la détection de signaux non-Gaussiens noyés dans du bruit Gaussien [1] (applications en radar, sonar, ...etc).

La plupart des techniques développées à ce jour sont fondées sur les polyspectres (cf. [6, 2]) (on teste généralement que le bispectre ou le trispectre est "statistiquement" nul). Ces méthodes ne sont appropriées que si l'on dispose d'un grand nombre d'échantillons (la convergence des estimateurs polyspectraux étant "lente"). D'autre part, une approche temporelle a récemment été proposée dans le cadre des séries chronologiques corrélées, par Epps [3] ainsi que par Giannakis et Tsatsanis [5], basées respectivement sur la mesure de la fonction caractéristique et des cumulants d'ordre 3 ou 4 des processus. Cette dernière semble plus intéressante par rapport à l'approche fréquentielle (ou polyspectrale) lorsque les échantillons à tester sont relativement courts.

Nous présentons ainsi, dans cette contribution une approche temporelle où nous construisons une famille de tests statistiques basée sur la minimisation d'une forme quadratique faisant intervenir la différence entre la moyenne empirique et la moyenne d'ensemble de *fonctions non-linéaires* des données observées.

D'une façon générale, nous supposerons que le signal observé $\{X_t\}_{t\in\mathbb{Z}}$ est un processus stochastique, discret et stationnaire, de moyenne $E\{X_t\} = \mu$ et de fonction d'autocorrélation $r(\tau) = E\{(X_t - \mu)(X_{t-\tau} - \mu)\}$. Alors, le problème que nous nous posons se résume à tester l'hypothèse composite H_0:"*le processus* $\{X_t\}$ *est Gaussien*", sachant que les paramètres $\mu, r(0), r(1), r(2), \cdots$ sont inconnus.

Cette classe de tests peut être considérée comme l'équivalent des *'tests de conformité'* classiques (ou goodness-of-fit) étendus aux séries temporelles scalaires. Ces tests prennent en compte la loi de distribution conjointe des variables aléatoires $X_{t-\tau_1}, X_{t-\tau_2}, \cdots, X_{t-\tau_k}$ (voir [1, 8, 9]). L'originalité de l'approche proposée réside dans le fait que nous donnons un formalisme général des tests de conformité classiques (du type Chi-carré de Pearson, Skewness-Kurtosis ou autres), puisque le choix des transformations non-linéaires sur les données reste libre. Mise à part la stationnarité et une condition de régularité relativement faible concernant la densité spectrale, les tests que nous proposons ne requièrent aucune connaissance supplémentaire concernant le processus $\{X_t\}$.

2 Présentation du formalisme des tests de Gaussianité

Nous proposons dans cette étude une famille de tests statistiques permettant de valider l'hypothèse de *normalité* d'un processus stationnaire et réel. Les tests de normalité prendront en compte la loi de distribution conjointe d'un ensemble fini de variables aléatoires $(X_t, X_{t-\tau_1}, \cdots, X_{t-\tau_k})$ prises à différents instants (successifs ou pas).

2.1 Principe

Nous construisons le processus à N dimensions $\mathbf{Z}(t) = \{Z_1(t), \cdots, Z_N(t)\}^t$, défini à partir de N transformations non-linéaires effectuées sur le processus $\{X_t\}$.

$$Z_k(t) = g_k(X_t, X_{t-\tau_1^k}, \cdots, X_{t-\tau_{i(k)}^k}) \ , \quad k = 1, \cdots, N \quad (1)$$

où, $\Omega = \{\tau_j^k;\ 1 \le k \le N \text{ et } 1 \le j \le i(k)\}$ est un ensemble fini de retard et $g_k(.)$ sont certaines fonctions non-linéaires.

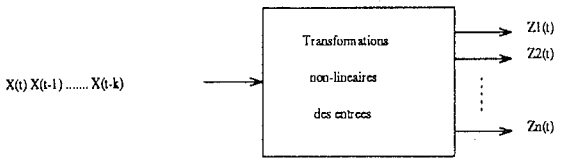

Figure 1: Schéma du principe

Il est clair de voir que sous l'hypothèse de base H_0, la loi de distribution du vecteur $\mathbf{Z}(t)$ dépend uniquement d'un ensemble fini de paramètres noté $\theta = (\mu, r(\tau_i^k - \tau_j^l);\ \tau_i^k, \tau_j^l \in \Omega)$ qui paramétrise la distribution conjointe du vecteur $\{X_0, X_\tau\ ,\ \tau \in \Omega\}$.

2.2 Loi des grands nombres et Théorème de la limite centrale

Il est important de remarquer que le processus $\{X_t\}_{t\in\mathbb{Z}}$ à étudier étant supposé stationnaire, alors le vecteur aléatoire transformé $\mathbf{Z}(t)$) est aussi stationnaire.

A ce niveau, la question importante à laquelle il nous est à répondre, est de savoir sous quelles conditions concernant le processus $\{X_t\}$ et ses transformations $g_k(.)$, le processus aléatoire vectoriel $\mathbf{Z}(t)$ est

- (i) régulier (c.a.d possédant une densité spectrale absolument continue sur $[-\pi, +\pi[$).
- (ii) vérifie une version du "théorème de la Limite Centrale".

2.2.1 Propriétés de mélange pour le processus $\{X_t\}$

Rosenblatt [10] et Ibragimov [7] ont défini la régularité généralisée d'un processus en introduisant la propriété de *mélange*. Deux définitions différentes du mélange ont été proposées:

- (i) $\sup_{A\in\mathcal{M}_{-\infty}^k,\ B\in\mathcal{M}_{k+n}^{+\infty}} |P(AB)-P(A)P(B)| \leq \alpha(n)$ où $\alpha(n) \to 0$ quand $n \to \infty$. $\mathcal{M}_a^b$ représente la σ-algèbre générée par les évènements du type $\{(X_{i_1}, \cdots, X_{i_k}) \in E\}$ où $a \leq i_1 < i_2 < \cdots < i_k \leq b$ et E un *Borélien* de dimension k.
- (ii) $\sum_{\tau_1,\cdots,\tau_k\in\mathbb{Z}} |cum^{(k)}(\tau_1, \cdots, \tau_k)| < +\infty$ où k est l'ordre du cumulant calculé à partir de $\{X_t\}$.

Ainsi, si nous faisons l'hypothèse que le processus $\{X_t\}$ est stationnaire, réel et Gaussien, il devient possible de connaître un ordre de grandeur du coefficient de mélange $\alpha(n)$.

Théorème 1 *(Ibragimov et Rozanov [11] pp.67)*
Si la fonction de distribution spectrale de $\{X_t\}$ est absolument continue et que la densité spectrales $f(\lambda)$ est de la forme $|P(e^{i\lambda})|^2 w(\lambda)$ où $P(z)$ est un polynôme dont les zéros sont sur le cercle unité $|z| = 1$ et $w(\lambda)$ bornée non nulle, r fois différentiable et pour laquelle la dérivée d'ordre r satisfait une condition de Hölder d'ordre β; alors le processus $\{X_t\}$ est fortement mélangeant et le coefficient de mélange $\alpha(n)$ est de l'ordre de $O(n^{-r-\beta})$.

2.2.2 Régularité de $\{\mathbf{Z}(t)\}$ et Loi limite

Supposons que $\{X_t\}$ est un processus *fortement mélangeant* (vérifiant le théoreme 1) et que $\alpha(n)$ est son coefficient de mélange. Considérons de plus, que nous observons un échantillon de taille T de $\{X_t\}$ que nous notons $\{X_1, X_2, \cdots, X_T\}$. Définissons alors, la moyenne empirique sur les échantillons transformés

$$S_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{Z}(t) \qquad (2)$$

où $\mathbf{Z}(t)$ est déterminé par la relation 1. Nous montrons alors, les résultats importants suivants.

Théorème 2 *Si les conditions suivantes sont vérifiées*

1. *$\{X_t\}$ est un processus stationnaire et Gaussien fortement mélangeant*
2. *Les fonctions non-linéaires $g_k(.)$ sont choisies de façon à ce que pour certaines valeurs $\delta > 0$, on ait $E_\theta\{|Z_k(t)|^{2+\delta}\} < \infty$ pour $k = 1, \cdots, N$ (bornitude des moments)*

alors,

$$(i) \quad S_T \to_{p.s} S(\theta) = E_\theta\{\mathbf{Z}(t)\}$$

$$(ii) \quad \Sigma(\theta) = Cov_\theta(\mathbf{Z}(t)) + 2\sum_{\tau=1}^{\infty} E_\theta\{(\mathbf{Z}(t) - S_\theta)(\mathbf{Z}(t-\tau) - S_\theta)^t\} < \infty$$

$$(iii) \quad \sqrt{T}\ (S_T - S(\theta)) \to_{\mathcal{L}} \mathcal{N}(0, \Sigma(\theta))$$

où θ est l'ensemble des paramètres caractérisant la loi de distribution du vecteur $\mathbf{Z}(t)$ sous l'hypothèse H_0.

Démonstration: Nous ne donnerons pas le détail de la démonstration, par contre nous en citerons l'idée principale. Il suffit d'appliquer le théorème 1 et de vérifier pour chaque composante $Z_k(t)$ $(k = 1, \cdots, N)$ le théorème 2.1 donné par Ibragimov [7] *pp.*368. ■

Les résultats (ii) et (iii) premettent respectivement de prouver la régularité du processus transformé $\mathbf{Z}(t)$ et d'en vérifier le théorème de la Limite Centrale.

2.3 Construction des statistiques de test

Nous construisons la famille de tests statistiques en définissant la forme quadratique $Q_T(\theta)$:

$$Q_T(\theta) = T(S_T - S(\theta))^t\ \Sigma(\theta)^{-1}\ (S_T - S(\theta)) \qquad (3)$$

Le théorème 2 nous permet de connaitre la *distribution asymptotique* de la statistique de test 3 sous l'hypothèse de base H_0. Alors, sous les conditions énumérées dans ce théorème et en supposant que les paramètrs θ sont connus, $Q_T(\theta)$ est asymptotiquement distribuée selon une loi du χ^2 à N degrés de liberté. Ce résultat fondamental est à la base de la construction de la classe de tests de conformité que nous proposons.

3 Mise en oeuvre des procédures de test

L'implémentation de ce type de tests d'hypothèse composite pose deux problèmes, car dans la pratique,

1. La matrice de covariance $\Sigma(\theta)$ n'est pas connue.
2. Le vecteur θ des paramètres de la distribution de $\{\mathbf{Z}(t)\}$ est aussi inconnu.

Nous présentons alors dans les paragraphes qui suivent, des solutions qui permettent de construire des estimateurs consistants de ces deux quantités.

3.1 Estimation de la matrice de covariance de la statistique de test

Nous énonçons le théorème suivant.

Théorème 3 *Soit $\hat{\Sigma}_T$ un estimateur consistant de $\Sigma(\theta)$, c.à.d que $\hat{\Sigma}_T \rightarrow_{P_\theta} \Sigma(\theta)$. La forme quadratique $\tilde{Q}_T(\theta) = T(S_T - S(\theta))^t \hat{\Sigma}_T^{-1} (S_T - S(\theta))$ est asymptotiquement distribuée selon la loi du χ^2 avec N degrés de liberté. La distribution asymptotique de $Q_T(\theta)$ n'est pas affectée si l'on remplace $\Sigma(\theta)$ par $\hat{\Sigma}_T$, sous la condition que $\hat{\Sigma}_T \rightarrow_{P_\theta} \Sigma(\theta)$.*

Il est facile de voir que sous l'hypothèse H_0:

$$\lim_{T \rightarrow +\infty} TE_\theta\{(S_T - S(\theta))(S_T - S(\theta))^t\} = \Sigma(\theta) = 2\pi f_Z(0) \tag{4}$$

où $f_Z(\lambda)$ ($\lambda \in [-\pi, +\pi[$) est la *matrice de densité spectrale* du processus régulier $\{\mathbf{Z}(t)\}$ (cf. théorème 2).

Alors pour construire un estimateur consistant de $\Sigma(\theta)$, il suffit d'estimer la matrice de densité spectrale de $\{\mathbf{Z}(t)\}$ à la fréquence nulle. Nous pouvons utiliser selon notre choix:

- des périodogrammes lissés autour de la fréquence zéro (voir méthode proposée dans [8, 9]),
- des périodogrammes moyennés dans le domaine temporel,
- ou des méthodes quelconques procurant des estimateurs consistants de $f_Z(0)$.

3.2 Estimation des paramètres θ sous H_0

Puisque le vecteur des paramètres θ est inconnu, nous devons donc l'estimer. Soit $\hat{\theta}_T$ un tel estimateur; le test statistique utilisé en pratique est $\tilde{Q}_T(\hat{\theta}_T)$. Ceci change clairement la distribution de la statistique. Ce dernier point s'explique par le fait que $S(\hat{\theta}_T)$ devient un vecteur aléatoire et il n'est pas évident que la distribution asymptotique de la statistique de test soit de la même forme que celle de $\tilde{Q}_T(\theta)$ lorsque θ est connu. Par contre, il devient possible de calculer la loi limite de la statistique $\tilde{Q}_T(\hat{\theta}_T)$ si l'on choisit en particulier comme estimateur consistant des paramètres inconnus, le *minimisateur consistant* en θ de la forme quadratique 3. Plus précisément:

Théorème 4 *Soit θ la vraie valeur des paramètres et supposons aussi les conditions suivantes réalisées.*

1. *La dimension N de $\mathbf{Z}(t)$ est strictement supérieure au nombre de paramètres inconnus: $N > card(\theta)$*
2. *La matrice de densité spectrale $f_Z(0)$ est définie positive pour α dans un voisinage de θ*
3. *La fonction $S(\alpha)$ est deux fois continument différentiable en la variable α dans un voisinage de θ,*
4. *Le Hessien de $\tilde{Q}_T(\alpha)$ au point θ converge en probabilité (sous P_θ) vers une matrice inversible $I_Q(\theta)$*

$$I_{\tilde{Q}}^{(T)}(\theta) \stackrel{\text{def}}{=} [\frac{\partial^2 \tilde{Q}_T(\alpha)}{\partial\alpha_i \partial\alpha_j}|_{\alpha=\theta}]_{1 \leq i,j \leq card(\theta)} \rightarrow_{P_\theta} I_Q(\theta) \tag{5}$$

Alors,

- *(i) Il existe un unique minimisateur* **consistant** *$\hat{\theta}_T$ de la forme quadratique $\alpha \rightarrow \tilde{Q}_T(\alpha)$.*

$$\hat{\theta}_T = Argmin_{\alpha \in V_\theta} \tilde{Q}_T(\alpha)$$

- *(ii) La statistique $\tilde{Q}_T(\hat{\theta}_T)$ converge asymptotiquement en loi vers une distribution du χ^2 à $N - card(\theta)$ degrés de liberté.*

Remarque: Il existe un voisinage V_θ de θ, tel que $\alpha \rightarrow \tilde{Q}_T(\alpha)$ admette un minimum unique dans V_θ.

Pour ne pas allourdir le texte, nous ne présonterons pas la démonstration de ce thérorème. L'application de ce résultat pour en dériver une statistique de test, nécessite la résolution d'un problème de minimisation multidimensionnelle; ce qui risque de poser de sérieux problèmes numériques (ex: pour le cas à 1 retard τ, on a $\theta = \{\mu, r(0), r(\tau)\}$, donc $N > 3$).

Nous évitons le problème de la minimisation multidimensionnelle en faisant un approximation de l'estimateur $\hat{\theta}_T$ par la méthode *Quasi-Newton* ou communément appelée *scoring*.

Théorème 5 *Soit $\overline{\theta}_T$ un estimateur consistant du paramètre θ. Définissons l'estimateur $\tilde{\theta}_T$,*

$$\tilde{\theta}_T = \overline{\theta}_T - I_{\tilde{Q}}^{(T)}(\alpha)^{-1} \frac{\partial \tilde{Q}_T(\alpha)}{\partial\alpha}|\alpha = \overline{\theta}_T \tag{6}$$

où $I_{\tilde{Q}}^{(T)}(\overline{\theta}_T)$ est le Hessien de $\tilde{Q}_T(\alpha)$ au point $\overline{\theta}_T$ d'après 5. Alors,

- *l'estimateur $\tilde{\theta}_T$ converge en P_θ-probabilité vers l'unique minimisateur consistant $\hat{\theta}_T$ de $\alpha \rightarrow \tilde{Q}_T(\alpha)$.*
- *$\tilde{Q}_T(\tilde{\theta}_T)$ a la même distribution asymptotique que $\tilde{Q}_T(\hat{\theta}_T)$.*

Cette dernière solution est utilisée en pratique, avec pour point initial $\overline{\theta}_T$ l'estimateur empirique des paramètres inconnus.

4 Présentation de deux tests de Gaussianité

4.1 Test SK basé sur la mesure du Skewness et Kurtosis

$$Z(t) = \{X_t, X_t^2, X_t^3, X_t^4, X_t X_{t-\tau}, X_t^2 X_{t-\tau}, X_t^3 X_{t-\tau}\}^t$$

et sous l'hypothèse nulle H_0,

$$S(\theta) = \{ \mu, r(0)+\mu^2, 3\mu r(0)+\mu^3, 3r(0)^2+6\mu^2 r(0)+\mu^4, r(\tau)+\mu^2, 2\mu r(\tau)+\mu r(0)+\mu^3, 3r(0)r(\tau)+3\mu^2 r(0)+3\mu^2 r(\tau)+\mu^4\}^t$$

où $\theta = (\mu, r(0), r(\tau))$.

Le test SK permet de tester la normalité de la distribution conjointe de X_t et $X_{t-\tau}$.

4.2 Test ECF basé sur la fonction caractéristique empirique

$$Z(t) = \{\exp(i\lambda_1^t \overline{\mathrm{X}}(t)), \cdots, \exp(i\lambda_N^t \overline{\mathrm{X}}(t))\}^t$$

où λ_k $(1 \leq k \leq N) \in \mathbb{R}^p - \{0\}$ et $\overline{\mathrm{X}}(t) = \{X_t, \cdots, X_{t-p+1}\}^t$.

Sous l'hypothèse H_0, nous avons:

$$S(\theta) = \{\exp(i\mu\lambda_1^t \mathbf{1} + \frac{1}{2}\lambda_1^t R \lambda_1), \cdots, \exp(i\mu\lambda_N^t \mathbf{1} + \frac{1}{2}\lambda_N^t R \lambda_N)\}^t$$

où R est la matrice de Toeplitz $(p \times p)$ construite à partir de $r(0), r(1), \cdots, r(p-1))$. Le vecteur des paramètres inconnus est: $\theta = (\mu, r(0), \cdots, r(p-1))$.

Le test ECF permet de tester la normalité de la distribution conjointe de $X_t, \cdots, X_{t-p+1}$.

5 Application à la détection d'un processus non-Gaussien

L'observation est modélisée par: $X_t = Y_t + kA_t$ où Y_t est un processus Gaussien centré et autorégressif d'ordre 1 (le paramètre AR du modèle est inconnu). A_t représente le processus additif; il est supposé être i.i.d et distribué selon une loi exponentielle symétrique de facteur d'échelle $\lambda = 1$. Nous faisons varier le paramtre k de façon à analyser la puissance des tests pour différentes valeurs du rapport Signal/Bruit (SNR $\rho = k^2 \frac{E\{A_t^2\}}{E\{Y_t^2\}}$). La probabilité de détection est évaluée pour une fausse alarme de 5%, 2048 échantillons et 300 tirages aléatoires indépendants pour chaque test (SK et ECF).

Paramètres utilisés:

1. Test SK: nous choisissons un seul retard ($\tau = 1$ et N=7).

2. Test ECF: nous prenon un seul retard ($p = 2$ and $N = 24$ points de mesure pour la fonction caractéristique dans $\mathbb{R}_+^2 - \{0\}$).

Dans ces deux cas de figure, nous testons la normalité de la distribution conjointe de $\{X_t, X_{t-1}\}$.

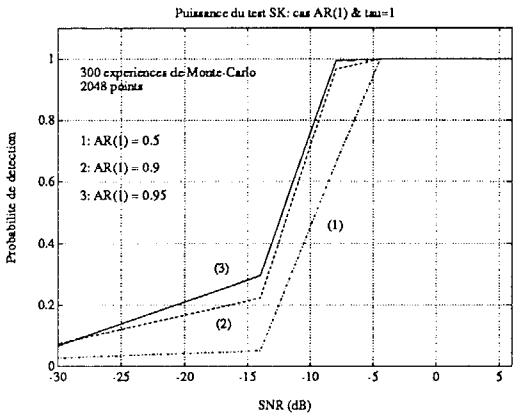

Figure 2: Test SK dans une situation de processus non-Gaussien additif.

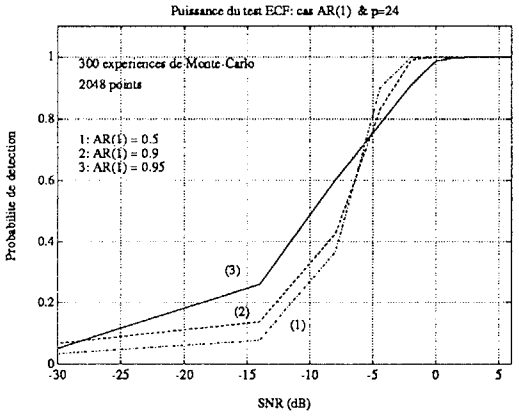

Figure 3: Test ECF dans une situation de processus non-Gaussien additif.

References

[1] S. Csörgo. Testing for Normality in Arbitrary Dimension. *The Annals of Stat.*, 14:708–723, 1986.

[2] J.W. Dalle Molle and M. Hinich. Cumulant Spectra-Based Tests for Detection of a Coherent Signal in Noise. In *Proc. of the Workshop on Higher-Order Statistics.*, pages 151–153. IEEE, 1991.

[3] T.W. Epps. Testing that a Stationary Time-Series is Gaussian. *The Annals of Stat.*, 15(4):1683–1698, 1987.

[4] T. Gasser. Goodness-of-Fit Tests for Correlated Data. *Biometrika*, 62:563–570, 1975.

[5] G.B. Giannakis and M.K. Tsatsanis. A Unifying Maximum-Likelihood View of Cumulant and Polyspectral Measures for Non-Gaussian Signal Classification and Estimation. *IEEE Tr. on IT*, 38(2):386–406, 1992.

[6] M.J. Hinich. Testing for Gaussianity and Linearity of a Stationary Time-Series. *Jour. of Time-Series Anal.*, 3(3):169–176, 1982.

[7] I.A. Ibragimov. Some Limit Theorems for Stationary Processes. *Theory Probab. Appl.*, 7:349–382, 1962.

[8] E. Moulines, K. Choukri, and M. Charbit. Testing that a Multivariate Stationary Time-Series is Gaussian. In *Proc. of the 6th SSAP Workshop on Stat. Sig. and Array Process.*, pages 185–188. IEEE, October 1992.

[9] E. Moulines, J.W. Dalle Molle, K. Choukri, and M. Charbit. Testing that a Stationary Time-Series is Gaussian: Time-domain vs. Frequency-domain Approaches. In *Proc. of IEEE Signal Process. Workshop on Higher-Order Statistics.*, pages 336–339. IEEE, June 1993.

[10] M. Rosenblatt. Asymptotic Normality, Strong Mixing and Spectral Density Estimates. *The Annal. of Probab.*, 12(4):1167–1180, 1984.

[11] M. Rosenblatt. *Stationary Processes and Random Fields.* Birkhauser–Verlag, 1985.