Un Test de Normalité pour les Processus Colorés Multivariés

Sara EL BOUCH, Olivier J. J. MICHEL, Pierre COMON*

Univ. Grenoble Alpes, CNRS, Gipsa-Lab
11 Rue des Mathématiques, BP 38400 Saint-Martin-d'Hères, France
sara.el-bouch@gipsa-lab.grenoble-inp.fr, olivier.michel@gipsa-lab.grenoble-inp.fr
pierre.comon@gipsa-lab.grenoble-inp.fr

Résumé – Nous nous intéressons à la détection d'événements rares, brefs et oscillants qui apparaissent comme non-gaussiens dans des données enregistrées simultanément sur un réseau de capteurs. L'outil choisi pour chiffrer l'écart à la gaussianité est le *kurtosis multivarié*. Nous utilisons une approche par fusion de détections binaires pour contrôler le taux d'erreur global. Le détecteur est opérationnel et nous étudions ses performances de détection sur des données réelles bruitées. Les résultats obtenus montrent que le test multivarié proposé présente non seulement un excellent pouvoir de détection pour de faibles RSB, mais encore conserve une bonne robustesse vis à vis des erreurs de modélisation.

Abstract – We are interested in the detection of rare, brief and oscillating events that appear as non-Gaussian in data recorded simultaneously on a sensor network. We concentrate on the *Multivariate Kurtosis* to measure the gap from Gaussianity. We use a binary detection fusion framework to control the overall error rate. The detector is operational in a real-time environment and we validate its detection performance on real noisy data. The results obtained imply that the test satisfies desirable requirements : it exhibits a good nobustness to model's misspecification and a high detection power even at low SNR.

1 Introduction

L'hypothèse de normalité des données est très souvent à la base des théories et développements algorithmiques en traitement du signal. Dans les cas où une telle hypothèse n'est pas satisfaite, il est à présent bien compris qu'une solution n'utilisant que des moments d'ordre deux ne peut être satisfaisante – un cas emblématique est celui de la séparation aveugle de sources [2]. De nombreuses procédures ont donc logiquement été développées pour tester la validité de cette hypothèse. Elles sont regroupées sous la dénomination de *tests de normalité*. La majorité de ces procédures testent la gaussianité d'échantillons (mono ou multivariés) *indépendants*; citons par example le test de Lilliefors [9] basé sur la fonction de répartition empirique, le test de Mardia [11] pour des échantillons multivariables, et le test de Shapiro-Wilk [13]. De plus rares travaux [5, 10, 12] ont été menés pour tester si un processus *coloré* est gaussien. Ce constat est exacerbé dans le cas d'un processus *vectoriel*, malgré la croissance actuelle de l'intérêt pour les séries temporelles multivariées issues de mesures sur réseaux de capteurs. C'est pourquoi nous avons développé un test de normalité *vectoriel* pour les processus multivariés qui relaxe l'hypothèse d'indépendance [3] : les observations sont identiquement distribuées mais non indépendantes temporellement. Nous ne faisons aucune hypothèses sur les corrélations instantanées spatiales entre les d composantes des observations. Contrairement aux méthodes fondées sur le polyspectre, comme le test d'Hinich [7], qui nécessitent un grand nombre d'échantillons, le test proposé fonctionne sur des données de taille moyenne et se distingue par son faible coût de calcul.

Le test développé est un test de normalité sans alternative. Sous l'hypothèse gaussienne $\mathcal{H}_0$, toutes les statistiques peuvent s'exprimer à l'aide des seuls moments d'ordre deux. La variance d'estimation de ces derniers est proportionnelle au temps caractéristique de corrélation des observations. Suivant l'approche décrite dans [8], un blanchiment partiel des données est donc systématiquement appliqué, par filtrage vectoriel auto-régressif d'ordre p :

$$\boldsymbol{x}(n,i) \approx \sum_{k=1}^{p} \boldsymbol{A}_{k,i}\boldsymbol{x}(n-k,i) + \boldsymbol{\epsilon}(k,i) \qquad (1)$$

où $\{\boldsymbol{x}(n,i) \in \mathbb{R}^d,\, n = 1,\ldots,N\}$ représente l'ensemble des observations menées sur le capteur ou le sous groupe de capteurs indéxé par i, $\boldsymbol{A}_{k,i}$ sont des matrices $d \times d$, et $\boldsymbol{\epsilon}(i) \overset{def}{=} [\boldsymbol{\epsilon}(1,i),\ldots,\boldsymbol{\epsilon}(N,i)]^T \underset{i.i.d}{\sim} \mathcal{N}(0,\boldsymbol{S}(i))$ sont les résidus de régression, gaussiens sous $\mathcal{H}_0$. En pratique, l'ordre p du filtre blanchisseur est inconnu et doit être estimé; le signal vectoriel observé ne suit pas exactement un modèle auto-regréssif: le blanchiment n'est donc que partiel et le développement d'un test de normalité dans le cas coloré prend tout son sens.

Le problème que nous nous posons se résume sous la forme de tests binaires multiples pour tout $1 \leq i \leq N_c$:

$$\mathcal{H}_0^{(i)} : \boldsymbol{\epsilon}(i) \underset{n.i.d}{\sim} \mathcal{N}(0,\boldsymbol{S}(i)) \quad versus \quad \bar{\mathcal{H}}_0^{(i)} \qquad (2)$$

où les résidus de chaque capteur sont distribués de manière identique mais non indépendante *(n.i.d)*. Illustrons cette problématique en considérant une application qui a une longue et

*Ce travail a été soutenu par la chaire MIAI "Environmental issues underground" de l'Institut MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

riche histoire en géophysique: la détection des tremblements sismiques. La loi de Gutenberg-Richter [6] dicte que le nombre cumulatif de tremblements de terre augmente exponentiellement avec la diminution de la magnitude. Ces événements de faible magnitude sont sévèrement contaminés par le bruit de fond. Pour les détecter, notre approche consiste à considérer que le bruit suit le modèle (1) sous hypothèse de base $\mathcal{H}_0$, et la détection des événements revient à chercher un brusque écart à la gaussianité des résidus.

Nous nous concentrons sur le *kurtosis multivarié* défini en (5), qui mesure l'aplatissement de la distribution comparativement à la loi gaussienne. Nous avons défini la loi de cette statistique dans [3] puis étudié sa performance sur des données synthétiques: copules *n.i.d* [3] et leur projections sur un espace arbitraire de faible dimension (1 ou 2) [4]. Dans cette communication, le test de normalité sera défini, et la méthode de détection expliquée, après discussion de la modélisation du bruit de fond par un filtre auto-régressif vectoriel. En mettant en œuvre des tests binaires multiples, nous validons la méthode de détection sur des données sismiques réelles enregistrées par un réseau de capteurs à trois axes.

2 Mise en oeuvre de la procédure du test

On dispose de N observations enregistrées simultanément par N_c capteurs, $\boldsymbol{x}(n,i) \in \mathbb{R}^d$ est la réponse de l'instrument de mesure à la $i^{\text{ème}}$ source observée. Nous rappelons que $\boldsymbol{x}(n,i)$ est supposé suivre un modèle auto-régressif vectoriel VAR(p) sous l'hypothèse de base $\mathcal{H}_0$. $\boldsymbol{S}(\tau,i) = \mathbb{E}\{\boldsymbol{\epsilon}(n,i)\boldsymbol{\epsilon}(n-\tau,i)^T\}$ désigne la fonction de covariance du processus d'innovation $\boldsymbol{\epsilon}(i)$ et on convient de noter $\boldsymbol{S}(0,i) \stackrel{def}{=} \boldsymbol{S}(i)$.

2.1 Estimation du modèle VAR(p)

Choix de l'ordre p. Pour ajuster un modèle auto-régressif vectoriel VAR(p), la première étape consiste à choisir l'ordre p. Nous considérons un ensemble de modèles VAR(p) et un candidat est sélectionné en minimisant un des critères d'information de type *Bayesian* (BIC):

$$BIC(p) = -2\log\hat{L}(\boldsymbol{A}_{k,i}, \boldsymbol{S}(i)) + log(N)d^2p \qquad (3)$$

$\hat{L}$ est la fonction de vraisemblance.

Estimation des coefficients $\boldsymbol{A}_{k,i}$. Une fois l'ordre choisi, pour estimer les coefficients du modèle, nous utilisons l'approche *des Moindes Carrés Récursifs* pour minimiser la fonction coût:

$$J(n) = \sum_{j=1}^{n} \lambda_1^{n-j}\boldsymbol{\epsilon}(j,i)^T\boldsymbol{\epsilon}(j,i) \qquad (4)$$

où λ_1 (typiquement $\lambda_1 = 0.99$) est un facteur d'oubli exponentiel qui accorde plus de poids aux données récentes permettant ainsi au modèle linéaire de s'adapter aux changements dans les statistiques des observations.

2.2 Présentation du Test de Gaussianité

Nous adoptons le *kurtosis multivarié* défini dans [11]:

$$\hat{B}_d(N,i) = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{\epsilon}(n,i)^T\widehat{\boldsymbol{S}}^{-1}\boldsymbol{\epsilon}(n,i)\right)^2 \qquad (5)$$

avec

$$\widehat{\boldsymbol{S}}(i) = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{\epsilon}(k,i)\boldsymbol{\epsilon}(k,i)^T \qquad (6)$$

Sous l'hypothèse de base $\mathcal{H}_0$, nous connaissons la *distribution asymptotique* de $\hat{B}_d(N,i)$: Elle est asymptotiquement distribuée selon une loi Gaussienne $\mathcal{N}(\mu,\sigma^2)$.

Dans le cas d'échantillons *i.i.d*, Mardia [11] a dérivé la moyenne et la variance de $\hat{B}_d(N,i)$. Nous avons étendu ce résultat dans le cas d'échantillons spatialement et temporellement dépendants [3].

2.3 Statistiques asymptotiques du test

Nous présentons brièvement les formules de la moyenne et la variance de la statistique du test.

2.3.1 Processus scalaire coloré ($d = 1$)

$$\mathbb{E}\{\hat{B}_1\} = 3 - \frac{6}{N} - \frac{12}{N^2}\sum_{\tau=1}^{N-1}(N-\tau)\frac{S(\tau)^2}{S^2} + o(\frac{1}{N}) \qquad (7)$$

$$\mathbb{V}\text{ar}\{\hat{B}_1\} = \frac{24}{N}\Big[1 + \frac{2}{N}\sum_{\tau=1}^{N-1}(N-\tau)\frac{S(\tau)^4}{S^4}\Big] + o(\frac{1}{N}) \qquad (8)$$

Dans le cas scalaire, la *dépendance temporelle* entre échantillons $x(t)$ et $x(t-\tau)$ est prise en compte dans la fonction d'auto-covariance $S(\tau)$.

2.3.2 Processus bivarié coloré ($d = 2$)

$$\mathbb{E}\{\hat{B}_2\} = 8 - \frac{16}{N} - \frac{4}{N^2}\sum_{\tau=1}^{N-1}\frac{(N-\tau)Q_1(\tau)}{(S_{11}S_{22} - S_{12}^2)^2} + o(\frac{1}{N}) \qquad (9)$$

$$\mathbb{V}\text{ar}\{\hat{B}_2\} = \frac{64}{N} + \frac{16}{N^2}\sum_{\tau=1}^{N-1}\frac{(N-\tau)Q_2(\tau)}{(S_{11}S_{22} - S_{12}^2)^4} + o(\frac{1}{N}) \qquad (10)$$

Dans le cas bivarié, $Q_1(\tau)$ et $Q_2(\tau)$ sont des combinaisons linéaires des fonctions d'auto-covariance et d'inter-covariance ($S_{11}(\tau), S_{12}(\tau)$ et $S_{22}(\tau)$). Le détail de ces formules peut être retrouvé ici [3].

En pratique, la fonction de covariance est estimée à l'aide des observations disponibles:

$$\boldsymbol{S}(\tau,i) = \frac{1}{N}\sum_{j=1}^{N-\tau}\boldsymbol{\epsilon}(j,i)\boldsymbol{\epsilon}(j+\tau,i)^T \qquad (11)$$

Maintenant que la statistique du test est bien définie sous l'hypothèse de base $\mathcal{H}_0$, nous avons accès aux p-valeurs définies pour chaque capteur i par:

$$p_{(i)} = 2(1 - \Phi(z(i))) \qquad (12)$$

où $z(i) = \frac{\hat{B}_d(N,i)-\mu}{\sigma}$, μ (7, 9) et σ^2 (8, 10) selon la valeur de d. Φ désigne la fonction de répartion de $\mathcal{N}(0,1)$.

Un seul paramètre permet d'ajuster la détection: *le seuil de signification* ou l'erreur de première espèce défini par la probabilité:

$$\alpha = \mathbb{P}(\text{décider } \bar{\mathcal{H}}_0 | \mathcal{H}_0 \text{ est vraie}) \tag{13}$$

2.4 Projections aléatoires et procédure B-H

Ayant seulement défini les statistiques du test dans le cas scalaire et bivarié, nous projetons aléatoirement N_p fois les résidus de chaque capteur sur un plan ($d = 2$) ou sur une direction ($d = 1$).

Nous avons $m = N_p \times N_c$ hypothèses $\mathcal{H}_0^{(i)}$, $i = 1, \ldots, m$ à tester. Si tous les tests sont seuillés avec α, le niveau de fausse alarme est contrôlée à $m\alpha$ et conduit à un grand nombre de fausses alarmes, ou *fausse découvertes*. Une première solution consiste à seuiller chacun des tests par α/m, cette correction est due à Bonferroni.

Nous choisissons la procédure proposée par Benjamini et Hocheborg (BH) [1] pour contrôler le taux de fausses découvertes (FDR): le taux de vraies $\mathcal{H}_0^{(i)}$ rejetées à tort parmi toutes les hypothèses rejetées. Soit $0 < \delta \leq 1$ un paramètre de contrôle du FDR:

1. Soient $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ les p-valeurs ordonnées.
2. Soit $k = \underset{i}{\text{argmin}}\left(p_{(i)} \leq \frac{i}{m}\delta\right)$
3. Rejet des hypothèses $\mathcal{H}_0^{(1)}, \mathcal{H}_0^{(2)}, \ldots, \mathcal{H}_0^{(k)}$

3 Application à la détection des signaux sismiques

L'observation est modélisée par une contamination additive:

$$\boldsymbol{x}(n,i) = \text{Diag}\{\boldsymbol{a}\}\, \boldsymbol{s}(n,i) + \boldsymbol{b}(n,i) \tag{14}$$

$\boldsymbol{s}(n,i) \in \mathbb{R}^3$ est le signal sismique *réel* enregistré par le capteur i parmi $N_c = 8$ capteurs; $\boldsymbol{a} \in \mathbb{R}^3$ est un paramètre qui permet de régler le Rapport Signal/Bruit; le processus $\boldsymbol{b}$ est soit:

1. *un bruit synthétique:* un processus centré gaussien filtré par un passe-bas VAR(5), qui peut par exemple modéliser le bruit de mesure décorrélé d'un capteur à l'autre.
2. *un bruit réel:* un extrait de bruit sismique ambient *i.e.* une référence bruit seul (cf. Fig. 3).

Dans le premier cas (synthétique), nous étudions l'impact du choix de l'ordre du filtre blanchisseur RLS sur les performances du test pour $p = 5$ (l'ordre correct, cf. Fig. 1) et $p = 2$ (ordre mal estimé cf. Fig. 2).

Les performances de la procédure détaillée dans la section 2.4 sont estimées sur $M = 1000$ tirages aléatoires des deux statistiques de test $\hat{B}_1(N)$ et $\hat{B}_2(N)$. Nous comparons le pouvoir empirique calculé comme: $\frac{\#Rejections}{M}$ pour plusieurs valeurs de Rapport Signal/Bruit (RSB).

- Le paramètre de contrôle du FDR est $\delta = 5\%$.
- La taille des échantillons est $N = 1000$. Le nombre de capteurs à trois axes est $N_c = 8$.
- Les résidus obtenus par blanchiment de chaque réponse du capteur sont projetés aléatoirement $N_p = 5$ fois sur un plan qui passe par l'origine de l'espace initial (tridimensionnel). Pour comparer avec les performances de $\hat{B}_1$, les mêmes résidus sont projetés $N_p = 5$ fois sur une direction arbitraire (aléatoire).

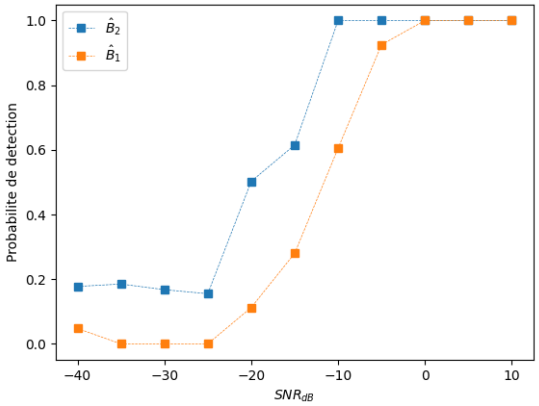

FIG. 1: Probabilité de détection dans une situation de processus Gaussien additif auto-régressif d'orde 5 blanchi par un VAR(5)

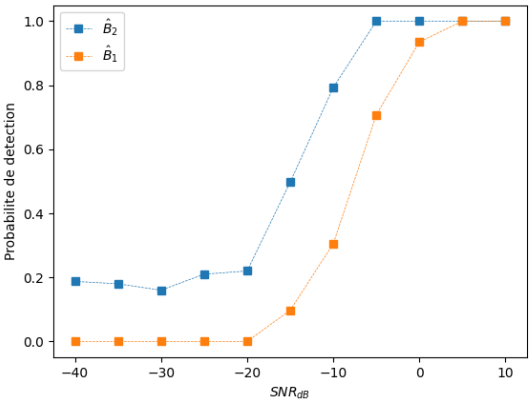

FIG. 2: Probabilité de détection dans une situation de processus Gaussien additif auto-régressif d'orde 5 blanchi par un VAR(2)

Enfin, notre souci étant de proposer un détecteur opérationnel en environnement réel, nous effectuons une dernière simulation sur la réponse d'un seul instrument de mesure. Les observations bruitées sont filtrés par un filtre auto-régressif vectoriel d'ordre 15. Les résidus sont ensuite projetés sur un plan arbitraire (aléatoire). Enfin, nous estimons de façon récursive la statistique du test et son seuil:

$$\hat{B}_2(n) = \lambda_2 \hat{B}_2(n-1) + (1-\lambda_2)\left(\boldsymbol{\epsilon}(n)^T \widehat{\boldsymbol{S}}^{-1} \boldsymbol{\epsilon}(n)\right)^2 \tag{15}$$

Le résultat de détection est présenté dans la Fig. 4.

- Pour toutes les simulations, le détecteur qui prend en compte la corrélation spatiale et temporelle $\hat{B}_2$ a un meilleur pouvoir de détection que son équivalent scalaire $\hat{B}_1$. Et ce, pour de très faibles RSB.

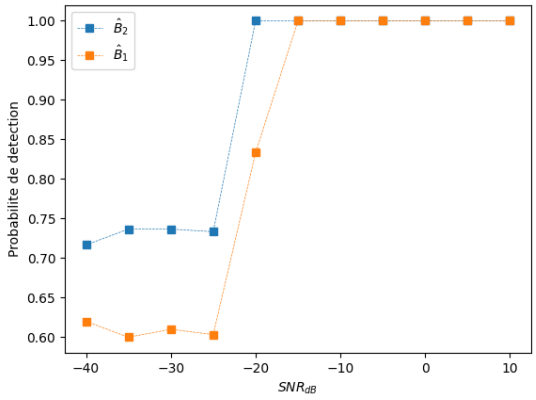

FIG. 3: Probabilité de détection dans une situation de processus extraits de données réelles blanchi par un VAR(15)

- Quand l'ordre du filtre de pré-blanchiment est mal estimé, la performance du détecteur basé sur $\hat{B}_1$ se dégrade. Les résultats du détecteur $\hat{B}_2$ sont bons pour un RSB ≥ -15dB.
- Dans le cadre d'une application sur des données réelles (Fig. 3), le détecteur $\hat{B}_2$ a un très bon pouvoir de détection ($\geq 70\%$) même pour de très faibles RSB.
- Dans la Fig. 4, l'arrivée d'une onde P se traduit par un pic dans la statistique du test.

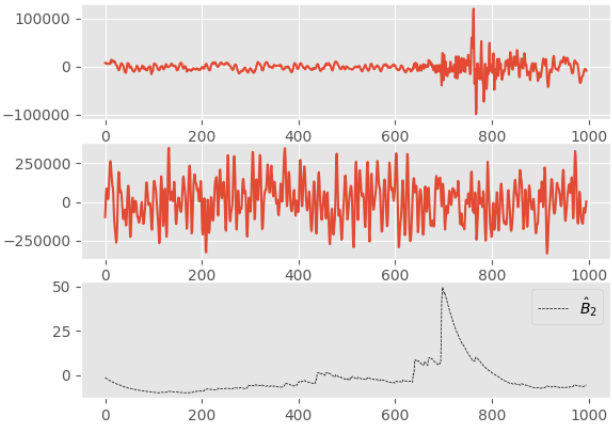

FIG. 4: Haut: composante (horizontale) d'un signal sismique; Milieu: observation du bruit additionné au signal sismique (RSB$= -5dB$); Bas: fonction de détection obtenue après projection bidimensionnelle des résidus de filtrage VAR.

4 Conclusion

Le blanchiment partiel des données par filtrage VAR améliore les performances des détecteurs. Toutefois, ignorer la couleur demeurant après filtrage entraîne un biais parfois prohibitif. Nous avons montré que la prise en compte de cette couleur dans le calcul du niveau du test conduit à de très bonnes performances, aussi bien sur signaux synthétiques que sur signaux réels, lorsque le test est mis en œuvre sur des projections aléatoires bivariées.

Ici le test de normalité était construit en fusionnant plusieurs détecteurs binaires délocalisés. Une perspective serait d'effectuer des projections d'une observation conjointe de tous les capteurs (grande dimension) sur des sous-espaces aléatoires de dimension 2; cette approche nécessiterait le transfert de beaucoup de données vers le centre de fusion.

References

[1] Y. Benjamini and Y. Hochberg. "Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing". In: *J. Royal Statist. Soc., Series B* 57 (Nov. 1995), pp. 289–300.

[2] P. Comon and C. Jutten, eds. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Oxford, Burlington: Academic Press, 2010.

[3] S. ElBouch, O. Michel, and P. Comon. "A Normality Test for Multivariate Dependent Samples". hal-03344745. Mar. 2022.

[4] S. ElBouch, O. Michel, and P. Comon. "Joint Normality Test Via Two-Dimensional Projection". In: *ICASSP*. hal-03369151. Singapore, May 2022.

[5] T. W. Epps. "Testing that a stationary time series is Gaussian". In: *The Annals of Statistics* 15.4 (1987), pp. 1683–1698.

[6] G Gutenberg and CF Richter. "Seismicity of the earth and associated phenomena, Howard Tatel". In: *Journal of Geophysical Research* 55 (1950), p. 97.

[7] M. Hinich. "Testing for Gaussianity and Linearity of a Stationary Time Series". In: *Jour. Time Series Analysis* 3.3 (1982), pp. 169–176.

[8] C. M. Jarque and A. K. Bera. "A test for normality of observations and regression residuals". In: *Int. Stat. Review* (1987), pp. 163–172.

[9] H. W. Lilliefors. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown". In: *J. Am. stat. Assoc.* 62.318 (1967), pp. 399–402.

[10] I. N. Lobato and C. Velasco. "A Simple Test of Normality for Time Series". In: *Econometric Theory* 20.4 (2004), pp. 671–689.

[11] K. V. Mardia. "Measures of Multivariate skewness and kurtosis with applications". In: *Biometrika* 57 (1970), pp. 519–530.

[12] A. Nieto-Reyes, J. A. Cuesta-Albertos, and F. Gamboa. "A random-projection based test of Gaussianity for stationary processes". In: *Computational Statistics & Data Analysis* 75 (2014), pp. 124–141.

[13] S. S. Shapiro and M. B. Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3/4 (1965), pp. 591–611.