Estimation consistante de la dimension minimale des représentations d'état de séries temporelles multivariables de grandes dimensions

Daria Tieplova, Philippe Loubaton

Estimation consistante de la dimension minimale des représentations d'état de séries temporelles multivariables de grandes dimensions

Daria TIEPLOVA[1], Philippe LOUBATON[2]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong,
Run Run Shaw Building, Pokfulam Road, Hong Kong

[2]Laboratoire d'Informatique Gaspard Monge, UMR 8049, Université Gustave Eiffel,
5 Bd. Descartes, Cité Descartes, 77454 Marne la Vallée Cedex 2 , France
daria.tieplova@u-pem.fr, philippe.loubaton@u-pem.fr

Résumé – Cet article est consacré à l'estimation de la dimension minimale P des représentations d'état d'une série temporelle multivariable $(y_n)_{n\in\mathbb{Z}}$ de dimension M définie comme une version bruitée (le bruit est gaussien et blanc temporellement) d'un signal utile $(u_n)_{n\in\mathbb{Z}}$ dont la densité spectrale est rationnelle et de petit rang. Cette étude est menée dans le régime des grandes dimensions dans lequel le nombre d'échantillons disponibles N et la dimension des observations M sont grands et du même ordre de grandeur. En utilisant des techniques de grandes matrices aléatoires, il est établi que l'analyse des plus grands coefficients de corrélation canonique estimés entre le passé et le futur de y permet d'évaluer de faccon consistante P si des conditions liées au rapport signal sur bruit et aux coefficients de corrélation canonique entre le passé et le futur du signal u sont vérifiées

Abstract – This paper is devoted to the estimation of the minimal dimension P of the state-space realizations of a high-dimensional time series $(y_n)_{n\in\mathbb{Z}}$, defined as a noisy version (the noise is white and Gaussian) of a useful signal $(u_n)_{n\in\mathbb{Z}}$ with low rank rational spectral density, in the high-dimensional asymptotic regime where the number of available samples N and the dimension of the time series M converge towards infinity at the same rate. Using large random matrix techniques, it is shown that the largest estimated canonical correlation coefficients between the past and the future of y allows to estimate P consistently, provided the signal to noise ratio and the true non zero canonical correlation coefficients between the past and the future of the useful signal u are large enough.

1 Introduction

L'estimation consistante de la dimension minimale P des représentations d'état d'un signal aléatoire stationnaire de dimension M $(y_n)_{n\in\mathbb{Z}}$ à partir des N observations $y_1, y_2, \ldots, y_N$ est un problème fondamental du domaine de l'analyse des séries temporelles. Un grand nombre de travaux y ont été consacrés dans le régime asymptotique traditionnel dans lequel $N \to +\infty$ et M fixe, qui, en pratique, fournit des résultats pertinents si le rapport $\frac{M}{N}$ est suffisamment petit (voir par exemple [7] et les références qui y sont citées). Cependant, du fait du développement des grands réseaux de capteurs, il devient courant d'être confronté au cas où M et N sont grands et du même ordre de grandeur, une situation que l'on modélise classiquement par un double régime asymptotique, qualifié de régime des grandes dimensions, dans lequel M et N tendent vers $+\infty$ de telle sorte que $\frac{M}{N}$ converge vers une constante non nulle.

L'objet de cet article est d'aborder l'estimation de P dans ce régime dans le cas où l'observation $(y_n)_{n\in\mathbb{Z}}$ se met sous la forme $y_n = u_n + v_n$, où $(v_n)_{n\in\mathbb{Z}}$ est un bruit blanc temporellement de matrice de covariance $\sigma^2 I_M$ où σ^2 est un scalaire inconnu, et où le "signal utile" u est une série temporelle dont la densité spectrale est rationnelle et est, pour toute fréquence, sauf peut-être un nombre fini d'entre elles, une matrice de rang $K < M$. Ainsi que cela est bien connu, cette condition est équivalente à l'existence d'une représentation d'état minimale de u sous la forme

$$x_{n+1} = Ax_n + Bi_n,\ u_n = Cx_n + Di_n \qquad (1.1)$$

où $(i_n)_{n\in\mathbb{Z}}$ est un bruit blanc de dimension K de matrice de covariance I_K, où A est une matrice dont le rayon spectral est strictement plus petit que 1, la minimalité de la réprésentation (1.1) signifiant que la dimension de l'état x est minimale. L'objet de la théorie de la réalisation stochastique ([7]) consiste à déterminer la dimension P et les paramètres (A, B, C, D) de l'ensemble des représentations minimales de u à partir de la donnée de la densité spectrale de u, ou de faccon équivalente, à partir de la fonction d'autocovariance R_u de u définie par $R_u(k) = \mathbb{E}(u_{n+k}u_n^*)$. En particulier, si on désigne par $\mathbf{u}_n^L$ le vecteur de dimension ML $\mathbf{u}_n^L = (u_{n-L}^T, \ldots, u_n^T)^T$, alors, l'identification de P repose sur le fait que pour tout entier $L \geq P$, le rang de la matrice d'autocovariance $R_{f|p,u}^L = \mathbb{E}(\mathbf{u}_{n+L}^L \mathbf{u}_n^{L*})$ entre le "futur de profondeur L" et le "passé de profondeur L" est égal à P. Du fait de la blancheur temporelle du bruit v, il est immédiat de vérifier que $R_y(k) = \mathbb{E}(y_{n+k}y_n^*) = R_u(k)$ pour tout $k \neq 0$ et que $R_{f|p,y}^L = \mathbb{E}(y_{n+L}^L y_n^{L*})$ coïncide avec $R_{f|p,u}^L$ (la matrice $R_{f|p,u}^L$ dépend de $(R_u(k))_{k=1,\ldots,L}$, et

pas de $R_u(0)$). Dans ces conditions, P est aussi égal au rang de $R^L_{f|p,y}$ pour tout entier $L \geq P$. Cette propriété est à la base des approches traditionnelles d'estimation de P à partir de la donnée de $y_1, \ldots, y_N$. Ainsi, P est souvent estimé par le nombre de valeurs singulières significatives de l'estimée empirique $\hat{R}^L_{f|p,y}$ de $R^L_{f|p,y}$ définie par

$$\hat{R}^L_{f|p,y} = \frac{Y_{f,N} Y^*_{p,N}}{N},$$

où $Y_{f,N}$ et $Y_{p,N}$ sont définies par [1]

$$Y_{p,N} = \begin{pmatrix} y_1 & y_2 & \cdots & y_{N-1} & y_N \\ y_2 & y_3 & \cdots & y_N & y_{N+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_{N+L-2} & y_{N+L-1} \end{pmatrix} \quad (1.2)$$

et

$$Y_{f,N} = \begin{pmatrix} y_{L+1} & y_{L+2} & \cdots & y_{N-1+L} & y_{N+L} \\ y_{L+2} & y_{L+3} & \cdots & y_{N+L} & y_{N+L+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{2L} & y_{2L+1} & \cdots & y_{N+2L-2} & y_{N+2L-1} \end{pmatrix}. \quad (1.3)$$

Une autre approche est basée sur le fait que si $L \geq P$, P est également égal au nombre de coefficients de corrélation canonique non nuls entre les espaces $\mathcal{Y}_{p,L}$ et $\mathcal{Y}_{f,L}$ engendrés par les composantes des vecteurs $\mathbf{y}^L_n$ et $\mathbf{y}^L_{n+L}$ respectivement. Rappelons que ces coefficients sont par définition les valeurs singulières de la matrice $(R^L_y)^{-1/2} R^L_{f|p,y} (R^L_y)^{-1/2}$, qui bien entendu, sont inférieures à 1. P est alors estimé comme le nombre de coefficients de corrélation canonique significatifs entre les espaces engendrés par les lignes des matrices $Y_{p,N}$ et $Y_{f,N}$, c'est-à-dire le nombre de valeurs singulières significatives de la matrice $(\hat{R}^L_{f,y})^{-1/2} \hat{R}^L_{f|p,y} (\hat{R}^L_{p,y})^{-1/2}$ où $\hat{R}^L_{f,y} = \frac{Y_{f,N} Y^*_{f,N}}{N}$ et $\hat{R}^L_{p,y} = \frac{Y_{p,N} Y^*_{p,N}}{N}$.

Dans le régime asymptotique usuel défini par $N \to +\infty$ et M, K, P, L fixes, $\|\hat{R}^L_{i,y} - R^L_y\| \to 0$ for $i = p, f$ et $\|\hat{R}^L_{f|p,y} - R^L_{f|p,y}\| \to 0$, de sorte que les approches précédentes permettent d'estimer P de faccon consistante. Néanmoins, si M est grand et que le nombre d'échantillons disponibles N n'est pas suffisamment grand, le rapport $\frac{ML}{N}$ peut s'avérer trop élevé pour que les matrices $\hat{R}^L_{f|p,y}$, et $\hat{R}^L_{i,y}$, $i = p, f$, soient suffisamment proches au sens de la norme spectrale des vraies matrices $R^L_{f|p,y}$ and R^L_y respectivement. Ce type de situation peut être modélisé par un régime asymptotique dans lequel M et N convergent vers $+\infty$ de telle sorte que $c_N = \frac{ML}{N}$ converge une constante non nulle c_*. Dans ce contexte, la matrice $\hat{R}^L_{f|p,y}$ n'est pas un estimateur consistant au sens de la norme spectrale de $R^{(L)}_{f|p,y}$, et les valeurs singulières de $\hat{R}^L_{f|p,y}$ n'ont a priori aucune raison de se comporter comme celles de $R^{(L)}_{f|p,y}$. Les mêmes conclusions peuvent être formulées pour les matrices $(\hat{R}^L_{f,y})^{-1/2} \hat{R}^L_{f|p,y} (\hat{R}^L_{p,y})^{-1/2}$ et $(R^L_y)^{-1/2} R^L_{f|p,y} (R^L_y)^{-1/2}$. Dans la suite, nous synthétisons les résultats de [10] qui permettent de préciser le comportement des plus grandes valeurs singulières de $\hat{R}^L_{f|p,y}$ et $(\hat{R}^L_{f,y})^{-1/2} \hat{R}^L_{f|p,y} (\hat{R}^L_{p,y})^{-1/2}$ lorsque $c_N \to c_* < 1$ tandis que les entiers K, P et L restent fixes. Ainsi que le lecteur pourra le constater en parcourant [10], les plus grandes valeurs singulières de $\hat{R}^L_{f|p,y}$ ne permettent pas d'estimer P de faccon consistante. Dans ces conditions, nous nous concentrons ici sur celles de $(\hat{R}^L_{f,y})^{-1/2} \hat{R}^L_{f|p,y} (\hat{R}^L_{p,y})^{-1/2}$, et établissons qu'elles permettent d'estimer P de faccon consistante sous réserve de quelques conditions signifiant que le rapport signal sur bruit et les coefficients de corrélation canonique entre les espaces $\mathcal{U}_{p,L}$ and $\mathcal{U}_{f,L}$ engendrés par les composantes et $\mathbf{u}^L_n$ et $\mathbf{u}^L_{n+L}$ sont suffisamment grands. Nous remarquons que dans [10], la matrice de covariance du bruit blanc v n'est pas nécessairement un multiple de l'identité. Comme ce contexte plus général induit un certain nombre de complications, nous préférons expliciter les résultats de [10] lorsque $\mathbb{E}(v_n v^*_n) = \sigma^2 I_M$. Enfin, nous mentionnons que les outils développés dans [10] sont valides lorsque le bruit v est complexe gaussien circulaire, hypothèse que nous supposerons donc ici. Il est plus que probable que les conclusions de [10] restent valides dans le cas gaussien réel, ou dans le cas complexe non circulaire.

Nous terminons cette introduction en mentionnant les travaux précédents connectés au problème étudié dans cet article. A notre connaissance, aucune référence passée ne s'est intéressée à l'estimation de P dans le contexte des grandes dimensions. Une partie significative des techniques d'inférence statistique en grandes dimensions qui ont été développées en l'absence d'hypothèse supplémentaire de type parcimonie est basée sur l'estimation de K et des éléments propres de la matrice de covariance $\mathbb{E}(u_n u^*_n)$ du signal utile u dans le cas où u_n se met sous la forme $u_n = H s_n$ où H est une matrice constante $M \times K$ et $(s_n)_{n \in \mathbb{Z}}$ représente une série temporelle de dimension K (voir par exemple [8], [5]), et est basée sur des résultats connectés aux modèles "Spike" de grandes matrices aléatoires définis comme la somme d'une grande matrice aléatoire à éléments i.i.d. avec une matrice déterministe dont le rang reste fixe quand les dimensions de la matrice augmente (voir par exemple [2]). Notons également que dans le même contexte, [6] estime K en étudiant les valeurs propres de $\mathbb{E}(u_n u^*_{n-1})$. Il convient également de mentionner divers travaux menés dans le domaine de l'économétrie relatifs aux modèles dynamiques factoriels généralisés ([4] par exemple), mais le rapport signal sur bruit quand M et N augmentent est un terme $O(1)$ alors qu'il est typiquement $O(\frac{1}{N})$ dans notre contexte. Nous renvoyons le lecteur à la discussion de la section II de [9] pour de plus amples détails. Nous mentionnons enfin que quelques articles ont aussi étudié le comportement des coefficients de corrélation canonique entre les espaces engendrés par les lignes

[1] $\hat{R}^L_{f|p,y}$ dépend aussi des échantillons $(y_n)_{n=N+1,\ldots,N+2L-1}$ mais nous négligeons cet effet de bord

de 2 grandes matrices aléatoires X_1 et X_2. Mais les propriétés de X_1 et X_2 sont beaucoup plus simples à gérer que celles des matrices Y_p et Y_f, qui ne sont pas indépendantes entre elles, et pas à éléments indépendants et identiquement distribués (i.i.d.). Ainsi, [12] considère deux matrices aléatoires gaussiennes indépendantes entre elles, et dont les éléments sont i.i.d. [13] a généralisé les résultats de [12] au cas où X_1 et X_2 sont non gaussiennes. Enfin, dans [1], X_1 et X_2 sont gaussiennes à éléments i.i.d., mais $\mathbb{E}(X_1 X_2^*)$ est non nulle, mais se réduit à une matrice de petit rang.

2 Hypothèses.

Dans le régime asymptotique que nous considérons, il faut interpréter $M = M(N)$ comme un entier dépendant de N. Par conséquent, il nous faut d'abord préciser comment les quantités générant u dans la réprésentation d'état (1.1) se comportent quand N varie:

Hypothèse 2.1. *Nous supposons que $(i_n)_{n\in\mathbb{Z}}$, les matrices A et B, et donc l'état markovien $(x_n)_{n\in\mathbb{Z}}$, ne dépendent pas de N. La matrice $M \times P$ $C = C_N$ et la matrice $M \times K$ $D = D_N$ sont supposées rester bornées quand N augmente. Enfin, nous supposons que le rang r de la matrice de covariance $R^L_{u,N}$ du vecteur aléatoire de dimension ML $\mathbf{u}^L_n$, qui vérifie $r \leq P + KL$, ne varie pas avec N pour tout N assez grand.*

Soit alors $R^L_{u,N} = \Theta_N \Delta^2_N \Theta^*_N$ la décomposition en éléments propres de $R^L_{u,N}$ où Δ^2_N est la matrice diagonale $r \times r$ des valeurs propres non nulles de $R^L_{u,N}$. Alors, nous supposons que

Hypothèse 2.2. *Les matrices $r \times r$ Δ_N et $\Theta^*_N R^L_{f|p,N} \Theta_N$ convergent vers des limites Δ_* and Γ_* respectivement. De plus, la matrice Δ_* vérifie $\Delta_* > 0$.*

Cette dernière hypothèse apparaît nécessaire afin de pouvoir mettre en évidence des résultats de convergence solides des plus grands coefficients de corrélation canoniques entre les espaces colonnes de $Y_{p,N}$ et $Y_{f,N}$, et est la contrepartie des hypothèse classiques formulées dans le contexte des modèles "Spike" qui postulent que les valeurs singulières de la matrice déterministe de rang fixe tendent vers une limite.

Remarque 2.1. *L'hypothèse 2.2 implique que lorsque $N \to +\infty$, la matrice $r \times r$ de rang P $\Delta_N^{-1} \Theta^*_N R^L_{f|p,N} \Theta_N \Delta_N^{-1}$, dont les valeurs singulières sont les les coefficients de corrélation canonique entre les espaces $\mathcal{U}_{p,N}$ et $\mathcal{U}_{f,N}$ engendrés par les composantes des vecteurs $\mathbf{u}^L_n$ et $\mathbf{u}^L_{n+L}$, converge vers la matrice de rang P $\Omega_* = \Delta_*^{-1} \Gamma_* \Delta_*^{-1}$. Ω_* vérifie bien entendu $\|\Omega_*\| \leq 1$.*

3 Principaux résultats

Nous allons étudier le carré des valeurs singulières de la matrice $\Sigma = (\hat{R}^L_{f,y})^{-1/2} \hat{R}^L_{f|p,y} (\hat{R}^L_{p,y})^{-1/2}$, c'est-à-dire les valeurs propres de la matrice $\Sigma\Sigma^*$. Pour $i = p, f$, nous désignons par $\Pi_{i,y}$ la matrice de projection orthogonale sur l'espace engendré par les lignes de la matrice $Y_{i,N}$ qui est donnée par

$$\Pi_{i,y} = \frac{Y_i^*}{\sqrt{N}} \left(\frac{Y_i Y_i^*}{N} \right)^{-1} \frac{Y_i}{\sqrt{N}} \tag{3.1}$$

Alors, à l'exception de la valeur propre 0, il est facile de vérifier que les valeurs propres de $\Sigma\Sigma^*$ coïncident avec les valeurs propres la matrice $\Pi_{p,y}\Pi_{f,y}$. Si l'on désigne par $\Pi_{i,v}$ la matrice obtenue à partir de $\Pi_{i,y}$ en remplaccant les observations $(y_n)_{n=1,\dots,N}$ par les $(v_n)_{n=1,\dots,N}$, notre approche repose sur les étapes suivantes:

- Etape 1. Etablir que la distribution empirique des valeurs propres de $\Pi_{p,v}\Pi_{f,v}$ converge presque surement vers une mesure de probabilité $\tilde{\nu}_*$ dont le support $\mathcal{S}_*$ est donné par $\mathcal{S}_* = [0, 4c_*(1-c_*)] \cup \{1\}\mathbf{1}_{c_*>1/2}$, et montrer que pour tout $\epsilon > 0$, presque surement, à partir d'un certain rang, toutes les valeurs propres de $\Pi_{p,v}\Pi_{f,v}$ appartiennent à $[0, 4c_*(1-c_*) + \epsilon] \cup \{1\}\mathbf{1}_{c_*>1/2}$, la valeur propre 1 étant de multiplicité $2ML - N$ si $c_* > 1/2$.
- Etape 2. Remarquer que pour $i = p, f$, $\Pi_{i,y}$ est une perturbation de rang fini de la matrice $\Pi_{i,v}$, ce qui implique que $\Pi_{p,y}\Pi_{f,y}$ est elle-même une perturbation de rang fini de $\Pi_{p,v}\Pi_{f,v}$, s'inspirer de l'approche développée dans l'étude des modèles Spike conventionnels ([2], [3]) pour évaluer le nombre de valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ s'échappant de $\mathcal{S}_*$, et déterminer les conditions sous lesquelles ce nombre est égal à P. Sous ces conditions, P peut-être estimé de faccon consistante par le nombre de valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ s'échappant de $\mathcal{S}_*$.

La mesure probabilité $\tilde{\nu}_*$ est caractérisée par sa transformée de Stieltjes $\tilde{t}_*(z)$ dont l'expression analytique peut être calculée simplement. Par ailleurs, $\tilde{\nu}_*$ n'est rien d'autre que le produit de convolution multiplicatif libre ([11]) de $c_*\delta_1 + (1-c_*)\delta_0$ avec elle-même, c'est-à-dire la limite de la distribution empirique des valeurs propres du produit de 2 projections orthogonales sur les espaces engendrés par les lignes de 2 matrices aléatoires $ML \times N$ à éléments gaussiens i.i.d. Il est donc assez remarquable de constater que, bien que les matrices $V_{p,N}$ et $V_{f,N}$ définies de la même faccon que $Y_{p,N}$ et $Y_{f,N}$, ne vérifient pas ces propriétés, la distribution empirique des valeurs propres de $\Pi_{p,v}\Pi_{f,v}$ converge quand même vers $\tilde{\nu}_*$.

On définit la matrice de rang P F_* donnée par

$$F_* = \Omega_*^*(I_r + \sigma^2\Delta_*^{-2})^{-1}\Omega_*(I_r + \sigma^2\Delta_*^{-2})^{-1} \tag{3.2}$$

Les valeurs propres de F_* sont réelles, appartiennent à $[0, 1)$, et $\|F_*\| < 1$. Alors, le résultat essentiel de cet article est le théorème suivant.

Théorème 3.1. • *La fonction f_* définie par*

$$f_*(x) = x \left(\frac{c_* \tilde{t}_*(x)}{(1-c_*)(\tilde{t}_*(x) + (1-c_*)/x)} \right)^2 \tag{3.3}$$

est strictement croissante sur $[4c_*(1-c_*), 1]$, vérifie $f_*(4c_*(1-c_*)) = \frac{c_*}{1-c_*}$, $f(1) = 1$ si $c_* < \frac{1}{2}$ et $f(1) = \left(\frac{c_*}{1-c_*}\right)^2$ si $c_* > \frac{1}{2}$.

- *Si $c_* \geq \frac{1}{2}$, l'équation*

$$\det\left(f_*(x)\, I_r - F_*\right) = 0 \tag{3.4}$$

n'a pas de solution dans $]4c_*(1-c_*), 1[$, et pour tout $\delta > 0$, presque surement, pour N assez grand, toutes les valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ appartiennent à $[0, 4c_*(1-c_*)+\delta]\cup[1-\delta, 1]$. Parmi les valeurs propres appartenant à $[1-\delta, 1]$, $2ML - N + \mathcal{O}(1)$ d'entre elles sont égales à 1, et $o(N)$ convergent vers 1.

- *Si $c_* < \frac{1}{2}$, l'équation (3.4) possède $0 \leq s \leq P$ solutions appartenant à $]4c_*(1-c_*), 1[$ où s est le nombre de valeurs propres (en prenant en compte les multiplicités) de F_* qui sont strictement plus grandes que $\frac{c_*}{1-c_*} < 1$. Si $\rho_{1,*}, \ldots, \rho_{s,*}$ sont les solutions correspondantes, alors, les s plus grandes valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ convergent presque partout vers $\rho_{1,*}, \ldots, \rho_{s,*}$, et pour tout $\delta > 0$, presque surement, pour N suffisamment grand, les $N - s$ valeurs propres restantes appartiennent à $[0, 4c_*(1-c_*) + \delta]$.*

Le Théorème 3.1 permet d'en déduire immédiatement des conditions sous lesquelles P peut être estimé par le nombre de valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ plus grandes que $4c_*(1-c_*)$.

Corollaire 3.1. *P coïncide avec le nombre de valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ strictement plus grandes que $4c_*(1-c_*)$ si et seulement si $c_* < \frac{1}{2}$ et si les P valeurs propres non nulles de F_* sont strictement plus grandes que $\frac{c_*}{1-c_*}$.*

En pratique, les P valeurs propres non nulles de F_* sont supérieures à $\frac{c_*}{1-c_*}$ si toutes les valeurs propres non nulles de Ω_* et celles de Δ_*^2/σ^2 sont suffisamment grandes. Compte tenu de la Remarque 2.1, ceci implique que les P coefficients de corrélation canonique non nuls entre les espaces $\mathcal{U}_{p,N}$ et $\mathcal{U}_{p,N}$ et le rapport signal sur bruit sont eux-mêmes suffisamment grands. Il est également intéressant de remarquer que si $c_* > 1/2$, $s = 0$, et qu'il apparaît impossible d'estimer P à partir des valeurs propres de $\Pi_{p,y}\Pi_{f,y}$. Compte tenu de la taille de l'article, nous renvoyons le lecteur à [10] pour la présentation de simulations numériques permettant d'étayer le Corollaire 3.1.

Remarque 3.1. *Nous concluons cet article en précisant les raisons pour lesquelles, assez curieusement, il est impossible d'estimer P de faccon consistante, à partir des plus grandes valeurs singulières de $\Upsilon_y = \hat{R}^L_{f|p,y}$, ou de faccon équivalente, à partir des plus grandes valeurs propres de $\Upsilon_y\Upsilon_y^*$. Il est montré dans [10] que presque surement, à partir d'un certain rang, toutes les valeurs propres de $\Upsilon_v\Upsilon_v^*$ se situent au voisinage d'un intervalle $[0, x_*]$, où x_* peut s'exprimer analytiquement en fonction de σ^2. Cependant, à la différence de la matrice $\Pi_{p,y}\Pi_{f,y}$, le nombre de valeurs propres de $\Upsilon_y\Upsilon_y^*$ qui s'échappent de $[0, x_*]$, peut prendre toute valeur entre 0 et $2r$, et apparaît en tous les cas presque indépendant de la valeur de P. Dans [10], pour tout $r \geq 1$, nous produisons des exemples dans lesquels $P = 1$, le rapport signal sur bruit et la plus grande valeur singulière non nulle de $R^L_{f|p,u}$ peuvent prendre des valeurs arbitrairement grandes, mais où $2r-1$ valeurs propres de $\Upsilon_{N,y}\Upsilon_{N,y}^*$ convergent vers des valeurs plus grandes que x_*. Il n'est pas si simple d'expliquer simplement pourquoi l'analyse des valeurs propres de $\Pi_{p,y}\Pi_{f,y}$ permet d'estimer P (sous réserve de conditions) et pas celles de $\Upsilon_y\Upsilon_y^*$. On peut cependant remarquer que l'analyse des coefficients de corrélation canonique entre le passé et le futur d'un signal à spectre rationnel a des vertus bien connues dans le domaine de l'identification des systèmes linéaires stochastiques ([7]). Nos résultats fournissent un nouvel exemple de la pertinence de cette approche.*

References

[1] Z. Bao, J. Hu, G. Pan, W. Zhou, "Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case", Ann. of Stat., vol. 47, no. 1, pp. 612-640, 2019.

[2] F. Benaych-Georges, R.R. Nadakuditi,"The singular values and vectors of low rank perturbations of large rectangular random matrices", J. Multivariate Anal., vol. 111 (2012), pp. 120-135.

[3] F. Chapon, R. Couillet, W. Hachem, X. Mestre, "The outliers among the singular values of large rectangular random matrices with additive fixed rank deformation" , Markov Processes and Related Fields, 20, 183–228 (2014).

[4] M. Forni, M. Hallin, M. Lippi, L. Reichlin, "The generalized dynamic factor model: Identification and estimation", Rev. Econ. Stat., vol. 82, no. 4, pp. 540-554, 2000.

[5] S. Kritchman, B. Nadler, "Non-parametric detection of the number of signals, hypothesis testing and random matrix theory, *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3930–3941, 2009.

[6] Z. Li, Q. Wang, J. Yao, "Identifying the number of factors from singular values of a large sample auto-covariance matrix", Ann. of Stat., vol. 45, no. 1, pp. 257-288, 2017.

[7] A. Lindquist, G. Picci, "Linear Stochastic Systems", Series in Contemporary Mathematics, Vol. 1, Springer, 2015.

[8] X. Mestre, M.A. Lagunas, "Modified subspace algorithms for DoA estimation with large arrays", *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598-614, February 2008.

[9] A. Rosuel, P. Vallet, P. Loubaton, X. Mestre, "On the detection of low-rank signal in the presence of spatially uncorrelated noise: a frequency domain approach", IEEE Transactions on Signal Processing, 69, art. no. 9496131, pp. 4458-4473, 2021.

[10] D. Tieplova, P. Loubaton, "On the largest singular values of certain large random matrices with application to the estimation of the minimal dimension of the state-space representations of high-dimensional time series", arXiv:2110.11627.

[11] Voiculescu, D. V., Dykema, K. J., Nica, A.: Free Random Variables. A Noncommutative

[12] K. W. Wachter, "The limiting empirical measure of multiple discriminant ratios", Ann. Statist., vol. 8, no. 5, pp. 937-957, 1980.

[13] Y. R. Yang, G. M. Pan, "The convergence of the empirical distribution of canonical correlation coefficients", Electron. J. Proba., 17: 1-13 (2012).