Error exponents for AR order testing

Stéphane Boucheron and Elisabeth Gassiat

***Abstract*— This paper is concerned with error exponents in testing problems raised by auto-regressive (AR) modeling. The tests to be considered are variants of generalized likelihood ratio testing corresponding to traditional approaches to auto-regressive moving-average (ARMA) modeling estimation. In several related problems like Markov order or hidden Markov model order estimation, optimal error exponents have been determined thanks to large deviations theory. AR order testing is specially challenging since the natural tests rely on quadratic forms of Gaussian processes. In sharp contrast with empirical measures of Markov chains, the large deviation principles satisfied by Gaussian quadratic forms do not always admit an information-theoretical representation. Despite this impediment, we prove the existence of non-trivial error exponents for Gaussian AR order testing. And furthermore, we exhibit situations where the exponents are optimal. These results are obtained by showing that the log-likelihood process indexed by AR models of a given order satisfy a large deviation principle upper-bound with a weakened information-theoretical representation.**

***Index Terms*— Time series; Error exponents; Large deviations; Gaussian processes; Order; Test; Levinson-Durbin**

I. Introduction

A. Nested composite hypothesis testing

THIS paper is concerned with composite hypothesis testing: a measurable space $(\Omega, \mathcal{A})$ and two sets of probability distributions $\mathcal{M}_0$ and $\mathcal{M}_1$ are given. In the sequel, we assume $\mathcal{M}_0 \subset \mathcal{M}_1$. A test is the indicator of a measurable set $K \subseteq \Omega$ called the detection region. The problem consists of choosing K so that if $P \in \mathcal{M}_0$, the level $P\{K\}$ is not too large while if $P \in \mathcal{M}_1 \setminus \mathcal{M}_0$, the power $P\{K\}$ should remain not too small.

If both $\mathcal{M}_0$ and $\mathcal{M}_1$ actually contain only one probability distribution, the hypothesis testing problem is said to be simple, and thanks to the Neyman-Pearson Lemma (see [48], [10]), test design is well-understood: for a given level, the most powerful test consists of comparing the likelihood ratio $\frac{P_1\{\mathbf{y}\}}{P_0\{\mathbf{y}\}}$ with a threshold.

When $\mathcal{M}_0$ and $\mathcal{M}_1$ are composite and nested, optimal test design and test analysis turn out to be much more complicated. As a matter of fact, most powerful tests may fail to exist. And rather than trying to construct a single test, it is common to resort to asymptotic analysis. A filtration $(\mathcal{A}_n)_{n\in\mathbb{N}}$ on Ω and a sequence of tests $(K_n)_{n\in\mathbb{N}}$ are considered where, for each n, K_n is $\mathcal{A}_n$ measurable. It is commonplace to search for sequences of tests with non-trivial asymptotic level $\sup_{P\in\mathcal{M}_0} \limsup_n \alpha_n(P) < 1$ with $\alpha_n(P) = P\{K_n\}$ and optimal asymptotic power $\inf_{P\in\mathcal{M}_1} \liminf_n 1-\beta_n(P)$ where $\beta_n(P) = 1 - P\{K_n\}$. A sequence of tests is said to be consistent if its asymptotic level is null while its asymptotic power is one.

In this paper, we focus on Gaussian auto-regressive processes. The measurable space $(\Omega, \mathcal{A})$ consists of $\mathbb{R}^{\mathbb{N}}$ provided with the cylindrical σ-algebra. Recall that a stationary Gaussian process $\dots Y_{-k}, \dots, Y_1, Y_2, \dots, Y_n \dots$ is an auto-regressive (AR) process of order r if and only if there exists a Gaussian independently identically distributed sequence $\dots X_{-k}, \dots, X_1, X_2, \dots, X_n \dots$ called the innovation process and a vector $\mathbf{a} \in \mathbb{R}^r$, where $(1, a_1, \dots, a_r)$ is called the prediction filter, such that for each $n \in \mathbb{Z}$:

$$Y_n + \sum_{i=1}^{r} a_i Y_{n-i} = X_n \,.$$

If a process is an AR process of order r but not an AR process of order $r-1$, it is said to be of order exactly r.

We are interested in testing the order of auto-regressive processes. The alternative hypotheses are:

$$H_0(r) \ : \ \text{“ the order of the auto-regressive process is } < r\text{”}$$

against

$$H_1(r) \ : \ \text{“ the order of the auto-regressive process is } \geq r\text{.”}$$

Testing the order of a process is related to order identification [41], [51], [44], [27], [26], [24], [21], [22] and thus to model selection [6], [50], [49], [7], [3], [45] . Note that testing the order of AR processes may be regarded as an instance of testing the order of Markov processes. In the finite alphabet setting, the latter problem has received distinguished attention during recent years [22], [24], [31]. Testing the order of an AR process may also be considered as a detection problem (see [38]).

B. Error exponents and Large Deviation Principles

As far as AR processes are concerned, consistent sequences of tests have been known for a while [36], [35], [37]. On the other hand, little seems to be known about the efficiency of AR testing procedures. In this paper, we adopt the error exponents perspective that has been used since the early days of information theory [23], [43], [31], [21], [40], [33].

A sequence of tests is said to achieve error exponent $E_0()$ (resp. $E_1()$) at $P \in \mathcal{M}_0$ (resp. $P \in \mathcal{M}_1$) if the corresponding sequence of level functions $\alpha_n()$ (resp. power functions $1 - \beta_n()$) satisfies

$$\liminf_n \frac{1}{n} \log \alpha_n(P) \leq -E_0(P) \ ,$$

respectively

$$\liminf_n \frac{1}{n} \log \beta_n(P) \leq -E_1(P) \ .$$

Manuscript received October 5, 2004; revised November xx, 2005. This work was supported by Network of Excellence PASCAL.

E. Gassiat is with the Department of Mathematics, Université Paris-Sud.

S. Boucheron is with the Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 7-Denis Diderot.

Error-exponents are non-trivial whenever they are positive.

Note that this notion of asymptotic efficiency is connected to other notions of asymptotic efficiency in classical statistics. For example, Bahadur efficiency provides a related but different approach to asymptotic efficiency [34], [48], [46], [39], both notions are usually investigated using large deviations methods [29].

The following definition gathers the basic concepts that are useful in large deviation theory (see [29] for details).

Definition 1 (Definition of LDP): A rate function on a topological space E is a function $I : E \mapsto [0, \infty]$ which is lower-semi-continuous. It is said to be a *good rate function* if its level sets $\{x \ : \ x \in E, I(x) \leq a\}$ are compact.
A sequence $(Z^n)_{n\geq 1}$ of random elements in E is said to satisfy the *large deviations principle* (LDP) with rate function I and linear speed if the corresponding sequence $(P_n)_{n\geq 1}$ of laws on E satisfies the following properties:

1) *Upper bound*: For any measurable closed subset C of E,
$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \leq - \inf_{x\in C} I(x) \, . \tag{1}$$
2) *Lower bound*: For any measurable open subset G of E,
$$\liminf_{n\to\infty} \frac{1}{n} \log P_n(G) \geq - \inf_{x\in G} I(x) \, . \tag{2}$$

Henceforth, if P and Q are two probability distributions such that the density of P with respect to Q, dP/dQ is well-defined, the relative entropy $K(Q \mid P)$ between P and Q is defined as the expected value under Q of the log-likelihood ratio $\log Q/P$: $K(Q \mid P) = \mathsf{E}_Q[\log Q/P]$, (see [20], [23], [29] for more material on this notion).

In this paper we will say that a large deviation principle admits an information-theoretical interpretation if, for any $x \in E$ such that $I(x) < \infty$, there exists a sequence $(Q_n)_n$ of probability distributions on E such that

1)
$$\lim_n \frac{1}{n} K(Q_n \mid P_n) = I(x) \, .$$
2) The sequence of image probability distributions $(Q_n \circ Z_n)_n$, converges weakly to δ_x, the probability mass function concentrated on x.

The information theoretical interpretation is often (but not always) at the core of Cramer's change of measure argument. The latter usually paves the way to the LDP lower bound (see [9], [42] for exceptions).

The Sanov Theorem on the large deviations of the empirical measure of an independently identically collected sample, is the prototype of a large deviation principle admitting an information-theoretical interpretation. [29].

C. Previous work

In most testing problems, provided there is a sufficient supply of limit theorems for log-likelihood ratios, upper-bounds on error-exponents can be obtained using an argument credited to Stein (see [29] and Section III below). Henceforth, those upper-bounds will be called Stein upper-bounds.

Checking whether the so-called Stein upper-bounds may be achieved or not is more difficult (this is also true for Bahadur efficiencies, see [46, discussion page 564]). In some simple but non-trivial situations like product distributions on finite sets, the Sanov Theorem [29] allows to check the optimality of generalized likelihood ratio testing (GLRT) (see [21] and references therein).

The possibility to check whether generalized likelihood ratio testing achieves the Stein upper-bound depends on the very nature of the large deviation principles (LDPs) satisfied by the relevant log-likelihood processes. The touchstone is whether the rate function of the LDP admits a full information-theoretical interpretation (as defined above) or not.

In the case of memoryless sources (see [21] and references therein) and the case of Markov order estimation [31], the fundamental role of the information-theoretical interpretation of the LDP rate function is hidden by type-theoretical arguments and by the fact that the existence of a finite-dimensional sufficient statistics makes the argument relatively straightforward. The importance of the information-theoretical interpretation of the LDP rate function (satisfied by the log-likelihood processes) becomes obvious when dealing with hidden Markov models. In the case of hidden Markov models on finite alphabets and finite hidden state spaces, it took nearly ten years to check that the non-trivial error exponents established in [43] actually match the upper bounds derived from the Stein argument [33]. When dealing with memoryless sources over general alphabets, not all models may be considered as exponential models (multinomial), and analyzing maximum likelihood estimation often has to rely on empirical processes techniques [47]. In that case, under weak integrability constraints on the likelihood process indexed by the models (weak Cramer conditions), the rate function of the LDP satisfied by the log-likelihood processes only admits partial information-theoretical interpretations (see [42]). Nevertheless, non-trivial error exponents are established by resorting to those partial information-theoretical representation properties of the LDP rate function [17] but the achievability of the Stein upper bounds on error exponents is still an open question.

D. Error exponents for stationary Gaussian hypotheses testing

When dealing with AR order testing, variants of generalized likelihood ratio testing may be investigated according to two directions. The first one attempts to take advantage of the fact that, just like in the case of Markov chains over finite alphabets, the parameters of the sampled AR processes remain identifiable when model dimension is over-estimated (see [15]). Moreover, there exists consistent estimators like the Yule-Walker estimator that rely on finite-dimensional statistics which large deviations properties can be investigated (see [9] for AR(1) processes).

The second line of investigation proceeds according to the approach described in [33]: analyze the large deviations properties of the log-likelihood processes indexed by the competing models. We will see at the end of Section II that the

two approaches may coincide. However, the ability to work with finite-dimensional (asymptotically consistent) statistics does not provide us with a safeguard.

Whatever the approach, the main difficulty consists of coping with the absence of an information-theoretical interpretation of the large deviation rate functions. This difficulty is due to the lack of steepness of the limiting logarithmic moment generating function of the log-likelihood vectors (see again [16], [9] and references therein for other examples of this phenomenon). Despite this impediment, we prove that when testing the order of auto-regressive processes, a variant of GLRT achieves non-trivial under-estimation exponents. This result is obtained by showing that even though the rate function governing the LDP of the log-likelihood process does not enjoy a full information-theoretical representation property, it does enjoy a partial information-theoretical representation property. This pattern of proof should be put into the perspective of [17].

E. Organization of the paper

The paper is organized as follows. Some concepts pertaining to the theory of Gaussian time series (like spectral density, prediction error, Levinson-Durbin recursion) are introduced in Section II. In Section III, limit theorems concerning log-likelihood ratios between stationary Gaussian processes are recalled. The interplay between prediction error and relative entropy allows to characterize the information divergence rate between an AR process of order exactly r and processes of lower order in Theorem 2. At the end of Section III, the Stein argument is carried out in order to check that there are no non-trivial over-estimation exponents in AR order testing, and to derive non-trivial upper-bounds on under-estimation exponents. The main results of this paper (non-triviality of under-estimation exponents) are stated in Section IV. It is also checked that in some non-trivial situations the Stein upper-bounds are achievable. The rest of the paper is devoted to the proof of the main results. LDPs for vectors of log-likelihoods are derived in Section V. In Section VI, we try to overcome the fact that, unlike the rate functions underlying the classical Sanov Theorem [29], the rate functions underlying the LDPs stated in Section V are not known to be representable as information divergence rates. In order to fill the gap, the rate function of the LDP exhibited in Section V is (weakly) related to information divergence rates through corollary 2. This relationship is then exploited in Section VII where the main result of the paper (Theorem 6) is finally proved.

II. Conventions

Background, motivations and a broader perspective on the material gathered in this Section can be found in [15] and [1].

As pointed out in the introduction, a Gaussian AR process is completely defined by the prediction filter and the variance of the innovation process.

Henceforth Θ^r denotes the (bounded) set of vectors $\mathbf{a} \in \mathbb{R}^r$ such that the polynomial $z \mapsto 1 + \sum_{i=1}^r a_i z^i$ has no roots inside the complex unit disc. The set AR(r) of AR processes of order r may be parametrized by pairs $(\sigma, \mathbf{a}) \in \mathbb{R}_+ \times \Theta^r$. Note that this is a full parametrization [10].

If $(Y_n)_{n\in\mathbb{Z}}$ is a stationary Gaussian process, then it is completely defined by its covariance sequence $(\gamma(k))_{k\in\mathbb{Z}}$ defined as $\gamma(k) = \mathbb{E}\left[Y_n Y_{n+k}\right]$. Under some mild summability conditions (that are always satisfied by AR processes), the covariance sequence defines a function on the torus $\mathbb{T} = [0, 2\pi]$ that captures many of the information-theoretical properties of the process.

Definition 2: [SPECTRAL DENSITY] The covariance sequence of a stationary Gaussian process is the Fourier series of the spectral density f of the process:

$$f(\omega) = \sum_{k\in\mathbb{Z}} \gamma(k) e^{\sqrt{-1}\omega k},$$

where ω belongs to the torus $\mathbb{T} = [0, 2\pi)$.

The spectral density of a stationary process is non negative on the torus $\mathbb{T}$. The spectral factorization theorem [15] asserts that f is the spectral density of a regular stationary process if and only if there exists a sequence (d_n) in $l_2(\mathbb{Z})$ such that

$$f(\omega) = \left| \sum_{n\in\mathbb{Z}} d_n e^{-\sqrt{-1}n\omega} \right|^2 .$$

The function f is the spectral density of a regular AR process of order r if and only if there exists an innovation variance σ^2, and a prediction filter $\mathbf{a} \in \mathbb{R}^r$ such that

$$f(\omega) = \frac{\sigma^2}{\left|1 + \sum_{i=1}^r a_i e^{-\sqrt{-1}\omega}\right|^2} . \tag{3}$$

Let $\mathcal{M}_r$ denote the set of spectral densities of form (3) where $\mathbf{a} \in \Theta^r$, and $\mathcal{F}_r$ its subset of spectral densities for which $\sigma = 1$. Note that a function $f \in \mathcal{M}_r$ belongs to $\mathcal{F}_r$ if and only if $\frac{1}{2\pi}\int_{\mathbb{T}} \log f = 0$.

A function f on $\mathbf{T}$ defines a sequence of $n \times n$ Toeplitz matrices $(T_n(f))_{n\in\mathbb{Z}}$

$$T_n(f)[i,j] = \frac{1}{2\pi}\int_{\mathbf{T}} f(\omega) e^{\sqrt{-1}(i-j)\omega} d\omega \text{ for } i,j \in \{0, n-1\} .$$

The function f is called the symbol of the Toeplitz matrix. If f is the spectral density of some stationary process, then $T_n(f)$ is the covariance matrix of the random vector $(Y_{m+1}, Y_{m+2}, \ldots, Y_{m+n})$ for any $m \in \mathbb{N}$.

In the sequel, if $\mathbf{A}$ denotes a matrix $\mathbf{A}^\dagger$ denotes the transposed matrix.

The log-likelihood of a sequence of observations $\mathbf{Y} = Y_1, \ldots, Y_n$ (interpreted as a column-vector) with respect to the spectral density $\sigma^2 f$ where $f \in \mathcal{F}_r$ will be denoted by $\ell_n\left(\sigma^2, f, \mathbf{Y}\right)$:

$$\ell_n\left(\sigma^2, f, \mathbf{Y}\right) = -\frac{1}{2n}\log\left(\sigma^{2n}\det(T_n(f))\right) - \frac{1}{2n\sigma^2}\mathbf{Y}^\dagger T_n^{-1}(f)\mathbf{Y}$$

A Theorem due to Szegö (see [11]) asserts that as n tends to infinity, $\frac{1}{n}\log\det(\det(T_n(f))) \to \frac{1}{2\pi}\int_{\mathbb{T}}\log f$ which is null if $f \in \mathcal{F}_r$. Another Theorem by Szegö motivates

the approximation of $T_n^{-1}(f)$ by $T_n\left(\frac{1}{f}\right)$. The quasi-Whittle criterion is now defined as:

$$\bar{\ell}_n\left(\sigma^2, f, \mathbf{Y}\right) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2n\sigma^2}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\,.$$

The following test will be considered throughout the paper.

Definition 3: [PENALIZED WHITTLE ORDER TESTING] Let $\text{pen}(n,p)$ be a sequence indexed by $\mathbb{N}\times\mathbb{N}$. Assume that $\text{pen}(n,p)$ is increasing with respect to the second variable. The penalized Whittle order test $\phi_n^{W,r}$ accepts $H_0(r)$ if and only if

$$\sup_{\sigma, f\in\mathcal{F}_p}\left\{\bar{\ell}_n(\sigma^2, f, \mathbf{Y}) - \text{pen}(n,p)\right\}$$

is maximum for some $p < r$. Let $\alpha_n^{W,r}(\cdot)$ be its level function and $1-\beta_n^{W,r}(\cdot)$ its power function.

At that point, it seems that we have to deal with $\mathbb{R}_+\times\Theta^r$ as a parameter space. As $\mathbb{R}_+$ is not bounded, this does not seem suitable for discretization of the parameter space. Fortunately, the following proposition shows that as far as order testing is concerned, we can disregard the variance of innovations σ^2 and focus on the prediction filter $\mathbf{a}$.

Proposition 1: [VARIANCE OF INNOVATION] The quasi-Whittle criterion is maximized by choosing $\sigma^2 = \inf_{f\in\mathcal{F}_r}\frac{\mathbf{Y}^\dagger T_n(1/f)\mathbf{Y}}{n}$, the maximal value of the criterion equals

$$-\frac{1}{2}\log\inf_{f\in\mathcal{F}_r}\frac{\mathbf{Y}^\dagger T_n(\frac{1}{f})\mathbf{Y}}{n} - \frac{1}{2}\,.$$

This prompts us to define modified criteria. In order to test whether the observed process is of order exactly r or $r-1$, we will compare

$$\inf_{f\in\mathcal{F}_r}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y} \text{ and } \inf_{f\in\mathcal{F}_{r-1}}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\ .$$

Finally, we will repeatedly need to understand how an AR process of order exactly r can be approximated by an AR process of order at most $r-1$. This will be facilitated by an algorithm that has proved to be of fundamental importance in AR modeling (see [15] for more details).

Definition 4: [INVERSE LEVINSON-DURBIN RECURSION] Let $(1,\mathbf{a})$ with $\mathbf{a}\in\mathbb{R}^r$ define the prediction filter of an AR Gaussian process with innovation variance σ^2. Then the inverse Levinson-Durbin recursion defines the innovation variance σ'^2 and the prediction filter $(1,\mathbf{b})$ with $\mathbf{b}\in\mathbb{R}^{r-1}$ of a regular AR process of order $r-1$ in the following way:

$$\begin{aligned} b_i &= \frac{a_i + a_r a_{r+1-i}}{1-a_r^2} \quad \text{for } i\in\{1,\ldots,r-1\},\\ \sigma'^2 &= \frac{\sigma^2}{1+a_r^2}\,. \end{aligned}$$

The Levinson-Durbin algorithm has not been designed in order to solve information-theoretical problems but rather in order to solve least-square prediction problems (its range of applications goes far beyond Gaussian processes). But in the Gaussian setting, least-square prediction and information-theoretical issues overlap. This will be illustrated in the following section.

III. INFORMATION DIVERGENCE RATES

Information divergence rates characterize the limiting behavior of log-likelihood ratio between process distributions. As the AR processes under consideration in this paper are stationary ergodic and even Markovian of some finite order, information divergence rates between AR processes are characterized by the Shannon-Breiman-McMillan Theorem (see [20], [5], [25]).

Theorem 1: [SHANNON-BREIMAN-MCMILLAN] If P and Q denote the distribution of two stationary centered Gaussian sequences with bounded spectral densities g and f that remain bounded away from 0, letting P^n and Q^n denote the image of P and Q by the first n coordinate projections, then the information divergence rate between P and Q, $\lim_n\frac{1}{n}K(\mathsf{P}^n\mid\mathsf{Q}^n)$ exists and is denoted by $K_\infty(g\mid f)$ or $K_\infty(\mathsf{P}\mid\mathsf{Q})$. The following holds P-almost-surely and also in $L_2(\mathsf{P})$:

$$\frac{1}{n}\log\frac{\mathsf{P}\{Y_{1:n}\}}{\mathsf{Q}\{Y_{1:n}\}} \to K_\infty(g\mid f).$$

The information divergence rate can be computed either from the spectral densities or from the prediction errors:

$$\begin{aligned} &K_\infty(g\mid f)\\ &= \frac{1}{4\pi}\int_{\mathbb{T}}\left(\frac{g}{f} - 1 - \log\frac{g}{f}\right)d\lambda \qquad (4)\\ &= \frac{1}{2}\left\{\log\frac{\sigma_f^2}{\sigma_g^2} - 1 + \mathsf{E}_g\left[\frac{(Y_0 - \mathsf{E}_f[Y_0\mid Y_{-\infty:-1}])^2}{\sigma_f^2}\right]\right\} \qquad (5)\end{aligned}$$

where σ_g^2 and σ_f^2 represent the variance of innovations associated with P and Q ($\log\sigma_f^2 = \frac{1}{2\pi}\int_{\mathbb{T}}\log f d\omega$ and $\log\sigma_g^2 = \frac{1}{2\pi}\int_{\mathbb{T}}\log g d\omega$).

Derivations of (4) can be found in [1],[25] or [15]. Equation (5) follows from the definition of Gaussian conditional expectations and from [5].

Equation (4) corresponds to the traditional description of the information divergence rate between two stationary Gaussian processes. Although, it does not look as explicit, Equation (5) emphasizes the already mentioned interplay between least-square prediction and information. It will prove very useful when characterizing the minimal information divergence rate between AR-processes of order $r-1$ and an AR-process of order exactly r.

The following class of functions will show up several times in the sequel.

Definition 5: [Definition of $\mathcal{H}$] Let $\mathcal{H}$ denote the set of non-negative self-conjugated polynomials h on the torus $\mathbb{T}$, that is of form

$$h(\omega) = a_0 + \sum_{i=1}^p a_i\left(e^{-\sqrt{-1}i\omega} + e^{\sqrt{-1}i\omega}\right) = a_0 + 2\sum_{i=1}^p a_i\cos(i\omega)$$

for real numbers $a_0, a_1, \ldots, a_p$ such that $a_0 + 2\sum_{i=1}^p a_i\cos(i\omega)\geq 0$ for all ω

Notice that if $h\in\mathcal{H}$ has degree p, then it may be written as the square of the modulus of a polynomial of degree p on

the torus, that is there exists real numbers $b_0, b_1, \ldots, b_p$ such that

$$h(\omega) = \left| b_0 + \sum_{i=1}^{p} b_i e^{-\sqrt{-1} i \omega} \right|^2 .$$

Indeed, h is a spectral density associated with a covariance sequence $(\gamma(k))_k$ which is zero for $|k| > p$, that is the covariance sequence of a moving-average process of order p. It is also the inverse of the spectral density of an AR process.

Corollary 1: Let g denote the spectral density of a stationary regular Gaussian process. If $1/f \in \mathcal{H}$ and $K_\infty(f \mid g) < \infty$ then f defines a stable AR process.

The next Theorem identifies the minimal information divergence rate between $g \in \mathcal{M}_r$ and spectral densities from $\mathcal{M}_{r-1}$. The pivotal role of that kind of result in the analysis of composite hypothesis testing procedures is outlined in [41], [21]. Theorem 2 is an analog of similar Theorems from [41], [43] or Lemma 6 and 7 from [33] (see also [14]). But, thanks to the special relationship between log-likelihood and prediction error in the Gaussian setting, Theorem 2 provides the exact value of the infimum and the precise point where it is achieved in parameter space.

Theorem 2: [I-PROJECTION ON LOWER ORDER AR PROCESSES] If P is the distribution of a regular AR process of order exactly r with spectral density g, prediction filter $\mathbf{a}$ and innovation variance σ^2, then

$$\inf_{f \in \mathcal{M}_{r-1}} K_\infty(f \mid g) = \frac{1}{2} \log\left(1 + {a_r}^2\right) .$$

The infimum is achieved by a stable AR process of order $r-1$ for which the prediction filter and variance of innovations are obtained by the inverse Levinson-Durbin recursion.

Henceforth, we will overload the classical notations: the spectral density of the AR process of order $r-1$ that realizes the infimum in $\inf_{f \in \mathcal{M}_{r-1}} K_\infty(f \mid g)$ will be called the I-projection of g on the set $\mathcal{M}_{r-1}$, it will be denoted by $f(g)$.

Proof: Any spectral density of stable auto-regressive process $r-1$, may be defined by a prediction filter $\mathbf{b} \in \Theta^{r-1}$ and an innovation variance σ'^2.

$$\begin{aligned}
&\inf_{f \in \mathcal{M}_{r-1}} K_\infty(f \mid g) \\
&= \inf_{f \in \mathcal{M}_{r-1}} \left[\frac{1}{2} \mathsf{E}_f \left[\log \frac{\sigma^2}{\sigma'^2} \right.\right. \\
&\qquad \left.\left. + \frac{\left(Y_0 - \mathsf{E}_g[Y_0 \mid Y_{-\infty:-1}]\right)^2}{\sigma^2} - 1 \right] \right] \\
&= \inf_{\mathbf{b} \in \Theta^{r-1}, \sigma'} \left[\frac{1}{2} \mathsf{E}_{\mathbf{b},\sigma'} \left[\log \frac{\sigma^2}{\sigma'^2} \right.\right. \\
&\qquad \left.\left. + \frac{\left(X_0 - \sum_{i=1}^{r-1} (b_i - a_i) Y_{-i} + a_r Y_{-r}\right)^2}{\sigma^2} - 1 \right] \right] \\
&= \inf_{\mathbf{b} \in \Theta^{r-1}, \sigma'} \left[\frac{1}{2} \left[\log \frac{\sigma^2}{\sigma'^2} \right.\right. \\
&\qquad + \frac{\mathsf{E}_{\mathbf{b},\sigma'} \left[\left(\sum_{i=1}^{r-1} (b_i - a_i) Y_{-i} - a_r Y_{-r} \right)^2 \right]}{\sigma^2} \\
&\qquad \left.\left. + \frac{\sigma'^2}{\sigma^2} - 1 \right] \right] .
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathsf{E}_{\mathbf{b},\sigma'} \left[\left(\sum_{i=1}^{r-1} \frac{(b_i - a_i)}{a_r} Y_{-i} - Y_{-r} \right)^2 \right] \\
&\quad = \mathsf{E}_{\mathbf{b},\sigma'} \left[\left(\sum_{i=1}^{r-1} \frac{(b_i - a_i)}{a_r} Y_{-r+(r-i)} - Y_{-r} \right)^2 \right]
\end{aligned}$$

cannot be smaller than the backward one-step prediction error of order $r-1$ of the process defined by f. The latter is the one-step prediction error of the stationary process obtained by time reversing the process $(Y_n)_{n \in \mathbb{N}}$. As the time-reversed process has the same covariance structure, it also has the same spectral density as the initial process. Hence backward one-step prediction error of order $r-1$ equals σ'^2.

This lower bound is achieved if the filter $([(b_i - a_i)]/a_r)_{i=1,\ldots,r-1}$ coincides with the backward prediction filter associated with the spectral density f. The coefficients of the backward prediction filter coincide with the coefficients of the forward prediction filter. Hence the lower bound is achieved if and only if

$$b_i - a_i = a_r b_{r-i} \quad \text{for all } i, 1 \leq i < r ,$$

that, is for the result of the inverse Levinson recursion. Hence the infimum is achieved by choosing

$$\sigma'^2 = \frac{\sigma^2}{1 + a_r^2}$$

and it equals

$$\frac{1}{2} \log\left(1 + a_r^2\right) .$$

■

A similar theorem characterizes the information divergence rate between g and $\mathcal{M}_{r-1}$.

Theorem 3: [LOWER ORDER AR PROCESSES] Let g denote the spectral density of a regular Gaussian auto-regressive process of order r. Then

$$\inf_{f\in\mathcal{M}_{r-1}} K_\infty(g \mid f) = -\frac{1}{2}\log\left(1-a_r^2\right) \ ,$$

and the infimum is achieved by the spectral density defined which is defined by the prediction filter resulting from the inverse Levinson-Durbin recursion and with variance of innovations equal to the forward-error prediction of this last prediction filter.

The proof of Theorem 3 parallels the proof of Theorem 2 and is omitted.

The following Theorems provide upper-bounds on achievable error exponents. Their proof is similar to the proof of Stein's Lemma concerning the efficiency of likelihood ratio testing for simple hypotheses (see [29]) . For the sake of self-reference, it is included in the Appendix.

Theorem 4: [TRIVIALITY OF OVER-ESTIMATION EXPONENT] Let $\alpha_n^r(\cdot)$ denote the level function of a sequence ϕ_n^r of tests of $H_0(r)$ against $H_1(r)$. If the asymptotic power function is every-where positive in $\mathcal{M}_r$, then for any auto-regressive process of order $r-1$ and distribution P,

$$\lim_{n\to+\infty}\frac{1}{n}\log\alpha_n^r(P) = 0 \ .$$

The next theorem provides a challenging statement and will motivate the rest of the paper.

Theorem 5: [STEIN UNDER-ESTIMATION EXPONENT] Let $\alpha_n^r(\cdot)$ denote the level function of a sequence ϕ_n^r of tests of $H_0(r)$ against $H_1(r)$. Let $(1-\beta_n^r(\cdot))_n$ denote its power function. If the asymptotic level function is everywhere bounded away from 1 in $\mathcal{M}_{r-1}$, then for any auto-regressive process of order exactly r, distribution P and spectral density $g\in\mathcal{M}_r$,

$$\liminf_{n\to+\infty}\frac{1}{n}\log\beta_n^r(P) \geq -\inf_{f\in\mathcal{M}_{r-1}} K_\infty\left(f \mid g\right).$$

Remark: It should be noted that thanks to Theorem 2, the Stein under-estimation exponent is non-trivial and that it does not depend on the variance of innovations $\frac{1}{2\pi}\int_{\mathbb{T}} g d\omega$.

Remark: Theorem 5 helps us in understanding the difference between error exponents as used in information theory and Bahadur efficiencies used in mathematical statistics. Bahadur efficiency is best defined by considering tests that reject say $\mathcal{M}_0$ for large values of some statistic T_n. Assume $P\in\mathcal{M}_1\setminus\mathcal{M}_0$, and assume that on some sample $y_1,\ldots,y_n$ collected according to P, $T_n(y_1,\ldots,y_n)=t$. Define the "level attained" as

$$L_n = \sup_{P'\in\mathcal{M}_0} P'\{T_n > t\} \ .$$

The Bahadur slope at P (if it exists) is defined as the P-almost sure limit of $-2n^{-1}\log L_n$. If the tests consist of comparing the logarithm of the ratio of the maximum likelihoods in models $\mathcal{M}_0$ and $\mathcal{M}_1$ with thresholds, then P-almost surely, T_n converges to $\inf_{P'\in\mathcal{M}_0} K_\infty\left(P \mid P'\right)$.

The distinction between error exponents and Bahadur slopes is exemplified by the fact that the quantity $\inf_{P'\in\mathcal{M}_{r-1}} K_\infty\left(P \mid P'\right)$ that shows up in Theorem 3 coincides with the Stein upper-bound on the Bahadur slope of GLRT at $P\in\mathcal{M}_1$ (see [46, Theorem 8.2.9] and [48, Theorem 16.12]).

In [46], Taniguchi and Kakizawa characterize the Bahadur slopes of some testing procedures among stationary Gaussian processes. Their results (Theorems 8.2.14, 8.2.15, 8.2.16) concern models indexed by compact intervals on $\mathbb{R}$ and do not seem to be easily extensible to the order testing problem considered here.

Although the techniques used in this paper do not allow us to prove that the Stein under-estimation exponents are everywhere achievable, it is worth mentioning that for the order testing procedures under consideration, the under-estimation exponents do not depend on the variance of innovations.

Proposition 2: [SCALE-INVARIANCE OF EXPONENTS] Let $\mathbf{a}\in\Theta^r\setminus\Theta^{r-1}$ denote a prediction filter of order exactly r. For all $\sigma>0$, let P_σ denote the probability distribution of an AR process of order r parametrized by $(\sigma,\mathbf{a})$.

If, for any integer p, $\mathrm{pen}(n,p)/n$ tends to 0 as n tends to infinity, the under-estimation exponent of the penalized Whittle tests do not depend on the variance of innovations:

$$\limsup_{n\to+\infty}\frac{1}{n}\log\beta_n^{W,r}(P_\sigma). = \limsup_{n\to+\infty}\frac{1}{n}\log\beta_n^{W,r}(P_1).$$

The proof of this proposition is given in the Appendix.

At that point, it is relevant to provide an alternative view at the quasi-Whittle criterion. Minimizing $\mathbf{Y}^\dagger T_n(1/f)\mathbf{Y}$ over $f\in\mathcal{F}_r$ turns out to be equivalent to minimize the forward prediction error

$$\sum_{t=1}^{n}\left(Y_t+\sum_{i=1}^{r}a_iY_{t-i}\right)^2$$

with respect to $\mathbf{a}\in\Theta^r$, assuming that $Y_t=0$ for all $t\leq 0$. The solution of the latter problem is known as the Yule-Walker estimate of the prediction filter of order r on the sample $\mathbf{Y}$ (see [15]). It can be computed efficiently thanks to the (direct) Levinson-Durbin algorithm. Moreover, if a_r denotes the r-th coefficient of the prediction filter of order r output by the Levinson-Durbin algorithm on the data $\mathbf{Y}$, an interesting aspect of the analysis of the Levinson-Durbin algorithm is the following relation:

$$\frac{\inf_{f\in\mathcal{F}_r}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}}{\inf_{f\in\mathcal{F}_{r-1}}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}} = 1-a_r^2 \ .$$

Hence, comparing of Whittle approximations of log-likelihoods boils down to comparing the absolute value of r-th reflection coefficient a_r, with a threshold. This is all the more interesting as the first r reflection coefficients only depend on the first $r+1$ empirical correlations $\sum_{t=i+1}^{n}Y_tY_{t-i}$, that is on a finite dimensional-statistic.

However, the possibility to approximate GLRT while relying on finite dimensional statistics does not seem to be of great help as far as investigating error exponents is concerned

(see [30] for more background on the interplay between the availability of finite-dimensional sufficient statistics and error exponents).

IV. Main theorems

From now on, P is the distribution of an auto-regressive Gaussian process of order exactly r, and spectral density $g \in \mathcal{F}_r$. Our goal is to prove that at least $\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^{W,r}(P) > 0$ and whenever possible that $\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^{W,r}(P)$ can be compared with some information-theoretical quantities.

Recall that $f(g)$ denotes the spectral density of the I-projection of g on the set $\mathcal{M}_{r-1}$ of spectral densities of AR processes of order $r-1$.

The following subset of $\mathcal{M}_r$ shows up in the analysis of the under-estimation exponent at g.

Definition 6: Let $F(g)$ be defined as the subset of spectral densities f of AR(r) processes such that for all n sufficiently large

$$T_n\left(1/f\right) + T_n^{-1}\left(g\right) - T_n\left(1/g\right)$$

is positive definite.

As $T_n^{-1}\left(g\right) - T_n\left(1/g\right)$ is non-positive (see Proposition 5), $F(g)$ defines a non-trivial subset of $\mathcal{M}_r$. In some special cases the triviality or non-triviality of $F(g)$ may be checked thoroughly. For example, if g is the spectral density of an AR(1) process, then $f(g)$ is the spectral density of an AR(0), and computations relying on Lemma 4, allow to check that $f(g) \in F$. As soon as we deal with processes of order 2, things get more complicated, as demonstrated by the following proposition.

Proposition 3: Let g be the spectral density of an auto-regressive process of order 2, with prediction filter (a_1, a_2). Then

$$f(g) \in F(g) \Longleftrightarrow \left(1+a_2\right)^2 > a_1^2\left(1+a_2^2\right) .$$

The proof of Proposition 3 is given in the Appendix.

In the sequel, as g remains a fixed element of $\mathcal{F}_r$, we will often omit to make the dependence on g explicit, and abbreviate $F(g)$ to F.

The main result of this paper is the following Theorem.

Theorem 6: [UNDER-ESTIMATION EXPONENT]
Let g denote the spectral density of an AR process of order exactly r and let $F(g)$ be defined as above. Let $L(g)$ be defined by

$$L(g) = \inf_{f\in\mathcal{M}_{r-1}} \left[K_\infty\left(f \mid g\right) - \inf_{h\in F(g)} K_\infty\left(f \mid h\right)\right] .$$

The followings hold:
a) $L(g) > 0$.
b) If, for any integer p, pen(n,p) tends to 0 as n tends to infinity, penalized Whittle tests have non trivial under-estimation exponents:

$$\limsup_{n\to+\infty} \frac{1}{n} \log \beta_n^{W,r}(P) \leq -L(g) .$$

The quantity $L(g)$ may or may not coincide with the Stein under-estimation exponent described in Theorem 5. For example, elementary computations reveal that $L(g)$ coincides with the Stein upper-bound when g is an AR(1) process. Note that, using the the connection between the Whittle test and tests concerning the Yule-Walker estimate pointed out after Proposition 2, and building on results from [9], it is possible to check directly that the Whittle test for AR(1) processes achieves the Stein under-exponent.

V. LDP for vectors of quadratic forms

In this section, $f_1, \ldots, f_d$ denote a collection of spectral densities of stable AR(r) processes ($f_i \in \mathcal{M}_r$). This collection defines a vector of quadratic forms $\left(\mathbf{Y}^\dagger T_n(1/f_i)\mathbf{Y}\right)_{i=1,d}$.

The basic concepts of large deviation theory were recalled in the introduction (see Subsection I-B).

Our goal is to prove a LDP upper bound for the tuple of quadratic forms $\left(\mathbf{Y}^\dagger T_n(1/f_i)\mathbf{Y}\right)_{i=1,d}$, when the time series $Y_1, \ldots, Y_n, \ldots$ is distributed as an AR(r) process with spectral density g. As under-estimation events correspond to large deviations of the log-likelihood process indexed by $\mathcal{F}_r$ (this qualitative statement will be turned into a formal one thanks to Definition 12 in Section VIII below) we aim at identifying the under-estimation exponents with the value of the rate function, or rather with a limit of values of the rate function evaluated at some well-chosen points. This goal will be achieved through Lemma 9.

The search for LDP for Toeplitz quadratic forms of Gaussian sequences has received deserved attention during the last fifteen years (see [2], [4], [28], [16], [9], [32] and early references in [18], [19], [13], [12]). Those papers, except [32], describe the large deviations of a single quadratic form while just assuming that the underlying time series is a regular stationary Gaussian process. The results described in those references need to be completed to fit our purposes.

The basic scheme of analysis in those papers remains quite stable: the logarithmic moment generating function of the quadratic form is related to the spectrum of a product of Toeplitz matrices. The limiting behavior of the spectrum is characterized using the asymptotic theory of Toeplitz matrices developed by Szegö and Widom (see [11] for a modern account). The main difficulty lies in the fact that understanding the limiting behavior of the spectrum of the Toeplitz matrices is not enough.

Definition 7: Λ_n is the logarithmic moment generating function of $\left(\mathbf{Y}^\dagger T_n\left(\frac{1}{f_1}\right)\mathbf{Y}, \ldots, \mathbf{Y}^\dagger T_n\left(\frac{1}{f_d}\right)\mathbf{Y}\right)$.
For any $\boldsymbol{\lambda} \in \mathbb{R}^d$,

$$\Lambda_n(\boldsymbol{\lambda}) = \log \mathbb{E}\left[\exp\left(\sum_{i=1}^d \lambda_i \mathbf{Y}^\dagger T_n\left(\frac{1}{f_i}\right)\mathbf{Y}\right)\right] .$$

For any $\boldsymbol{\lambda} \in \mathbb{R}^d$, $\bar{\Lambda}(\boldsymbol{\lambda})$,

$$\bar{\Lambda}(\boldsymbol{\lambda}) = \limsup_{n\to+\infty} \frac{1}{n}\Lambda_n(\boldsymbol{\lambda}).$$

The function $\bar{I}$ is the Fenchel-Legendre transform of $\bar{\Lambda}$: for any $\mathbf{y} \in \mathbb{R}^{\mathbf{d}}$,

$$\bar{I}(\mathbf{y}) = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{\mathbf{d}}} \left(\langle\boldsymbol{\lambda}, \mathbf{y}\rangle - \bar{\boldsymbol{\Lambda}}(\boldsymbol{\lambda})\right).$$

Definition 8: A point $\mathbf{y} \in \mathbb{R}^d$ is said to be an exposed point of $\bar{I}$ with exposing hyper-plane $\boldsymbol{\lambda} \in \mathbb{R}^d$ if and only if for all $\mathbf{y}' \in \mathbb{R}^d$:

$$\bar{I}(\mathbf{y}') > \bar{I}(\mathbf{y}) + \langle \boldsymbol{\lambda}, \mathbf{y}' - \mathbf{y} \rangle \, .$$

Note that by the very definition of convexity, the existence of a vector $\boldsymbol{\lambda}$ that satisfies $\bar{I}(\mathbf{y}') \geq \bar{I}(\mathbf{y}) + \langle \boldsymbol{\lambda}, \mathbf{y}' - \mathbf{y} \rangle$ can be taken for granted. The strict inequality makes the definition interesting. Note that as a point-wise limit of convex lower-semi-continuous functions, $\bar{\Lambda}$, is convex and lower-semi-continuous.

Theorem 7: Let $f_1, \ldots, f_d$ denote a collection of spectral densities of stable AR(r) processes.
a) The sequence of tuples of quadratic forms $\frac{1}{n}\left(\mathbf{Y}^\dagger T_n(1/f_i)\mathbf{Y}\right)_{i=1,d}$, satisfies a LDP upper-bound with good rate function $\bar{I}$.
b) The sequence of quadratic forms satisfies a LDP lower bound with rate function $\bar{I}(\mathbf{y})$ if $\mathbf{y}$ is an exposed point of $\bar{I}$ with exposing hyperplane $\boldsymbol{\lambda}$ such that $\lim_n \frac{1}{n}\Lambda_n(\boldsymbol{\lambda})$ exists and is finite, and infinity otherwise.

This Theorem is a direct application of Baldi's generalization of the Gärtner-Ellis Theorem (see [29], Chapter 4, Section 5). As stated, it is of little utility since we know next to nothing about $\bar{\Lambda}$. In the next section, we will check that $\bar{I}$ is non-trivial and that the set of exposed points is non-empty.

VI. Representation formulae of rate function

This section unveils the structure of $\bar{\Lambda}$ and $\bar{I}$. It prepares the derivation of the partial information-theoretical interpretation of the LDP upper bound stated in Theorem 7. Lemma 1 provides with a characterization of $\bar{\Lambda}(\boldsymbol{\lambda})$ when it is finite. This characterization only depends on the spectral densities $(f_i)_{i=1,d}$, and g, through the convex function Λ defined in Definition 9. The Fenchel-Legendre transform $I(\cdot)$ of this function $\Lambda(\cdot)$ is a leverage in the analysis of the rate function $\bar{I}$. As a matter of fact, for any $\mathbf{z} \in \mathbb{R}^d$, such that $I(\mathbf{z}) < \infty$, $I(\mathbf{z})$ may be identified as an information divergence rate between a carefully chosen AR-(r) process and the process with spectral density g, (see Lemma 2). Lemma 3 states that the supremum in the definition of $\bar{I}$ is achieved whenever $\bar{I}(\cdot)$ is finite. Moreover, an important consequence of Lemmas 2 and 3 is that although they differ, $\bar{I}(\cdot)$ and $I(\cdot)$ have the same effective domain (Corollary 2).

Definition 9: [Definition of Λ.] For $\boldsymbol{\lambda} \in \mathbb{R}^d$, let

$$\Lambda(\boldsymbol{\lambda}) = -\frac{1}{4\pi}\int_{\mathbf{T}} \log\left(1 - 2g\sum_{i=1}^{d}\frac{\lambda_i}{f_i}\right) d\omega \ .$$

The function Λ is strictly convex, lower-semi-continuous on

$$\mathcal{D}_\Lambda = \left\{ \boldsymbol{\lambda} \ : \ 1 - 2g\sum_i \frac{\lambda_i}{f_i} \text{ is non-negative on } \mathbf{T} \right\},$$

and finite on $\boldsymbol{\lambda}$'s in $\mathcal{D}_\Lambda$ such that $1/g - 2\sum_i \lambda_i/f_i$ is not the null function. This follows from the fact that for such $\boldsymbol{\lambda}$'s, $1/g - 2\sum_i \lambda_i/f_i$ belongs to the set $\mathcal{H}$ (see Definition 5), has isolated zeros, and $\log\omega$ is integrable at 0.

Lemma 1: [Characterization of $\bar{\Lambda}$.] $\bar{\Lambda}$ coincides with Λ on the set $\mathcal{D}_{\bar{\Lambda}}$ where it is finite. Moreover, if for any $\boldsymbol{\lambda}$ in $\mathcal{D}_\Lambda$, $f_{\boldsymbol{\lambda}}$ is defined as a function on $\mathbb{T}$ by

$$1/f_{\boldsymbol{\lambda}}(\omega) = 1/g(\omega) - 2\sum_{i=1}^{d}\frac{\lambda_i}{f_i(\omega)} \, ,$$

then $\boldsymbol{\lambda} \in \mathcal{D}_{\bar{\Lambda}}$ if and only if $f_{\boldsymbol{\lambda}} \in F(g)$.
Recall that $F(g)$ is defined at the beginning of Section IV.
Proof: From a well-known elementary result (the Cochran Theorem, see [16, Lemma 1]), it follows that for any integer n and any $\boldsymbol{\lambda} \in \mathbb{R}^d$,

$$\Lambda_n(\boldsymbol{\lambda}) = \begin{cases} -\frac{1}{2}\log\det\left(I_n - 2T_n(g)T_n\left(\sum_{i=1}^d \frac{\lambda_i}{f_i}\right)\right) \\ \text{if } \ T_n^{-1}(g) - T_n\left(\frac{1}{g}\right) + T_n\left(1/f_{\boldsymbol{\lambda}}\right) \\ \text{is definite positive.} \\ +\infty \quad \text{otherwise.} \end{cases}$$

Now

$$I_n - 2T_n(g)T_n\left(\sum_{i=1}^{d}\frac{\lambda_i}{f_i}\right) = \ T_n(g)\left(T_n\left(\frac{1}{f_{\boldsymbol{\lambda}}}\right) + T_n^{-1}(g) - T_n\left(\frac{1}{g}\right)\right) .$$

As $T_n(g)$ is definite positive for all n, if $f_{\boldsymbol{\lambda}} \notin F$, $\limsup_n \frac{1}{n}\Lambda_n(\boldsymbol{\lambda}) = \infty$.

On the other hand, if $f_{\boldsymbol{\lambda}} \in F$, we may use the following factorization:

$$I_n - 2T_n(g)T_n\left(\sum_{i=1}^{d}\frac{\lambda_i}{f_i}\right) = T_n(g)T_n\left(\frac{1}{f_{\boldsymbol{\lambda}}}\right) \left(I_n + T_n^{-1}\left(\frac{1}{f_{\boldsymbol{\lambda}}}\right)\left(T_n^{-1}(g) - T_n\left(\frac{1}{g}\right)\right)\right) .$$

The limits of $-\frac{1}{2n}\log\det(T_n(g))$ and $-\frac{1}{2n}\log\det(T_n(1/f_{\boldsymbol{\lambda}}))$ are readily identified as

$$-\frac{1}{4\pi}\int_{\mathbb{T}}\log g \, d\omega \quad \text{and} \quad -\frac{1}{4\pi}\int_{\mathbb{T}}\log\frac{1}{f_{\boldsymbol{\lambda}}}\, d\omega$$

thanks to Szegö's limit theorem (see [11, Theorem 5.2, page 124]). As

$$\Lambda(\boldsymbol{\lambda}) = -\frac{1}{4\pi}\int_{\mathbb{T}}\log\frac{g}{f_{\boldsymbol{\lambda}}}\, d\omega \, ,$$

it remains to check that

$$\lim_n \frac{1}{n}\log\det\left(I_n + T_n^{-1}\left(\frac{1}{f_{\boldsymbol{\lambda}}}\right)\left(T_n^{-1}(g) - T_n\left(\frac{1}{g}\right)\right)\right) = 0 \, .$$

Recall that $T_n\left(\frac{1}{g}\right) - T_n^{-1}(g)$ is the sum of two matrices of rank r, that it is non-negative and that the sum

of its eigenvalues is upper-bounded by $2r\left(1+\sum_{j=1}^{r}a_j^2\right)$ where $\mathbf{a}\in\Theta^r$ is the prediction filter associated with g (see Proposition 5 in the Appendix). This proves that $T_n^{-1}\left(1/f_{\boldsymbol{\lambda}}\right)\left(T_n^{-1}\left(g\right)-T_n\left(\frac{1}{g}\right)\right)$ has at most $2r$ non null (actually negative but larger than -1) eigenvalues, and their sum is uniformly lower-bounded. This is enough to prove that $\det\left(I_n+T_n^{-1}\left(1/f_{\boldsymbol{\lambda}}\right)\left(T_n^{-1}\left(g\right)-T_n\left(\frac{1}{g}\right)\right)\right)$ is smaller than 1 but remains bounded away from 0 and that the desired limit is actually null.

■

Definition 10: Let I be the Fenchel-Legendre transform of Λ : for any $\mathbf{y}\in\mathbb{R}^d$,

$$I(\mathbf{y})=\sup_{\boldsymbol{\lambda}\in\mathbb{R}^d}\left(\langle\boldsymbol{\lambda},\mathbf{y}\rangle-\Lambda(\boldsymbol{\lambda})\right).$$

Lemma 2: [REPRESENTATION FORMULA FOR I]
a) For all $\mathbf{y}\in\mathbb{R}^d$,

$$I(\mathbf{y})=\inf_{f\in\mathcal{M}_r}\left\{K_\infty\left(f\mid g\right)\ :\quad y_i=\frac{1}{2\pi}\int_{\mathbb{T}}\frac{f}{f_i}d\omega\text{ for all }i\right\},$$

where the right-hand is infinite when the infimum is taken over the empty set.
b) When $I(\mathbf{y})$ is finite, the infimum is attained at some $f_{\boldsymbol{\lambda}}$ and this $f_{\boldsymbol{\lambda}}$ is the spectral density of an AR(r) process:

$$\frac{1}{f_{\boldsymbol{\lambda}}}=\frac{1}{g}-2\sum_{i=1}^{d}\frac{\lambda_i}{f_i}\tag{6}$$

for some $\boldsymbol{\lambda}\in\mathcal{D}_\Lambda^\circ$ such that

$$I(\mathbf{y})=\langle\boldsymbol{\lambda},\mathbf{y}\rangle-\Lambda(\boldsymbol{\lambda}).$$

The proof of Lemma 2, consists of checking that the convex function Λ is essentially smooth according to definition 2.3.5 in [29].

Proof: [Lemma 2] Let us first check that if $\left(\boldsymbol{\lambda}^m\right)_{m\in\mathbb{N}}$ is a sequence of vectors from $\mathcal{D}_\Lambda^\circ$ that converges to $\boldsymbol{\lambda}\in\mathcal{D}_\Lambda\setminus\mathcal{D}_\Lambda^\circ$, then $\left(\|\bigtriangledown\Lambda\left(\boldsymbol{\lambda}^m\right)\|\right)_{m\in\mathbb{N}}$ converges to infinity.

For any $\boldsymbol{\lambda}^m\in\mathcal{D}_\Lambda^\circ$, by Lebesgue differentiation Theorem

$$\partial_i\Lambda_{|\boldsymbol{\lambda}^m}=\frac{1}{2\pi}\int_{\mathbb{T}}\frac{g/f_i}{1-2g\sum_{j=1}^{d}\lambda_j^m/f_j}\,\mathrm{d}\omega\ ,$$

and the following polynomial

$$\frac{1}{g}1-2\sum_{j=1}^{d}\lambda_j^m/f_j$$

is positive everywhere on $\mathbb{T}$.

Now as $\boldsymbol{\lambda}\in\mathcal{D}_\Lambda\setminus\mathcal{D}_\Lambda^\circ$, we have $\frac{1}{g}-2\sum_{j=1}^{d}\lambda_j/f_j\geq 0$ on $\mathbb{T}$. Either the polynomial $\frac{1}{g}-2\sum_{j=1}^{d}\lambda_j/f_j=0$ on $\mathbb{T}$, or it vanishes on at least one and at most finitely many points on $\mathbb{T}$. Hence in all cases

$$\frac{1}{2\pi}\int_{\mathbb{T}}\frac{g/f_i}{1-2g\sum_{j=1}^{d}\lambda_j/f_j}\,\mathrm{d}\omega=\infty\ .$$

By Fatou's Lemma

$$\begin{aligned}\liminf_m\partial_i\Lambda_{|\boldsymbol{\lambda}^m}&\geq\frac{1}{2\pi}\int_{\mathbb{T}}\liminf_m\frac{g/f_i}{1-2g\sum_{j=1}^{d}\lambda_j^m/f_j}\,\mathrm{d}\omega\\&=\frac{1}{2\pi}\int_{\mathbb{T}}\frac{g/f_i}{1-2g\sum_{j=1}^{d}\lambda_j/f_j}\,\mathrm{d}\omega=\infty\ .\end{aligned}$$

Thus $\left(\partial_i\Lambda_{|\boldsymbol{\lambda}^m}\right)_{m\in\mathbb{N}}$ tends to infinity for each $i\in\{1,\ldots,d\}$.

Let $\mathbf{y}$ be such that $I(\mathbf{y})<\infty$. Let us now show that there exists some $\boldsymbol{\lambda}$ such that

$$I(\mathbf{y})=\langle\boldsymbol{\lambda},\mathbf{y}\rangle-\Lambda\left(\boldsymbol{\lambda}\right).$$

There exists a sequence $\left(\boldsymbol{\lambda}^m\right)_{m\in\mathbb{N}}$ where $\boldsymbol{\lambda}\in\mathcal{D}_\Lambda^\circ$, such that

$$I(\mathbf{y})=\lim_m\langle\boldsymbol{\lambda}^m,\mathbf{y}\rangle-\Lambda\left(\boldsymbol{\lambda}^m\right).$$

If the sequence $\left(\boldsymbol{\lambda}^m\right)_m$ is bounded, it has an accumulation point $\boldsymbol{\lambda}$ in $\mathcal{D}_\Lambda$ since $\mathcal{D}_\Lambda$ is closed (see Definition 9). Then by lower-semi-continuity of Λ:

$$I(\mathbf{y})=\langle\boldsymbol{\lambda},\mathbf{y}\rangle-\Lambda\left(\boldsymbol{\lambda}\right).$$

Moreover $\boldsymbol{\lambda}\in\mathcal{D}_\Lambda^\circ$ and

$$\mathbf{y}=\bigtriangledown\Lambda_{|\boldsymbol{\lambda}}\,.$$

Let us check now that $\left(\boldsymbol{\lambda}^m\right)_m$ is indeed bounded. Assume the contrary for a while. If the sequence $\left(\boldsymbol{\lambda}^m\right)_m$ is not bounded, then the sequence $\left(\boldsymbol{\lambda}^m/\left\|\boldsymbol{\lambda}^m\right\|\right)_m$, has an accumulation point $\boldsymbol{\eta}$ on the unit sphere of $\mathbb{R}^d$. For each m such that $\|\boldsymbol{\lambda}^m\|\geq 1$, $\boldsymbol{\lambda}^m/\left\|\boldsymbol{\lambda}^{\boldsymbol{m}}\right\|\in\mathcal{D}_\Lambda^\circ$ since $\mathbf{0}\in\mathcal{D}_\Lambda^\circ$. Hence we may assume that $\boldsymbol{\eta}\in\mathcal{D}_\Lambda$.

For every $\omega\in\mathbb{T}$, $2g(\omega)\sum_{i=1}^{d}\boldsymbol{\eta}_i/f_i(\omega)<0$,

The sub-sequence $\left(\langle\boldsymbol{\lambda}^m,\mathbf{y}\rangle-\Lambda\left(\boldsymbol{\lambda}^m\right)\right)_m$ is equivalent to $\left(\frac{1}{2}\left(\|\boldsymbol{\lambda}^m\|\left\langle\boldsymbol{\eta},\mathbf{y}\right\rangle-\log\|\boldsymbol{\lambda}^m\|\right)\right)_m$. Hence, this subsequence converges to ∞, which contradicts the assumption $I(\mathbf{y})<\infty$.

Hence, for any $\mathbf{y}\in\mathbb{R}^d$, such that $I(\mathbf{y})<+\infty$, there exists some $\boldsymbol{\lambda}\in\mathcal{D}_\Lambda^\circ$ such that

$$I(\mathbf{y})=\langle\boldsymbol{\lambda},\mathbf{y}\rangle-\Lambda\left(\boldsymbol{\lambda}\right).$$

From the very definition of I and Λ, for every $\boldsymbol{\lambda}'\in\mathcal{D}_\Lambda$,

$$\Lambda\left(\boldsymbol{\lambda}'\right)\geq\Lambda\left(\boldsymbol{\lambda}\right)+\langle\boldsymbol{\lambda}'-\boldsymbol{\lambda},\mathbf{y}\rangle\,,$$

which entails $\mathbf{y}=\bigtriangledown\Lambda\left(\boldsymbol{\lambda}\right)$. Now define $f_{\boldsymbol{\lambda}}$ using equation (6), then $f_{\boldsymbol{\lambda}}\in\mathcal{M}_r$, for each $i\in\{1,\ldots,d\}$, $y_i=\frac{1}{2\pi}\int_{\mathbb{T}}f_{\boldsymbol{\lambda}}/f_i\,\mathrm{d}\omega$, and $I(\mathbf{y})=K_\infty\left(f_{\boldsymbol{\lambda}}\mid g\right)$. This proves that if $I(\mathbf{y})\leq\infty$, then

$$I(\mathbf{y})\geq\inf_{f\in\mathcal{M}_r}\left\{K_\infty\left(f\mid g\right)\ :\quad y_i=\frac{1}{2\pi}\int_{\mathbb{T}}\frac{f}{f_i}\,\mathrm{d}\omega\text{ for all }i\right\}.$$

We will now check that if the set $f,f\in\mathcal{M}_r$ with $y_i=\frac{1}{2\pi}\int_{\mathbb{T}}\frac{f}{f_i}\,\mathrm{d}\omega$, for $i\leq d$ is non-empty, then $I(\mathbf{y})$ is upper-bounded by information-divergence rates between g and elements of this set.

Let $f \in \mathcal{M}_r$ be such that for all $i \in \{1,\dots,d\}$, $y_i = \frac{1}{2\pi}\int_{\mathbb{T}} \frac{f}{f_i}\,\mathrm{d}\omega$, then for any $\boldsymbol{\lambda} \in \mathcal{D}_\Lambda$

$$\begin{aligned} & K_\infty(f \mid g) - \langle \boldsymbol{\lambda}, \mathbf{y}\rangle + \Lambda(\boldsymbol{\lambda}) \\ = \;& \frac{1}{4\pi}\int_{\mathbb{T}} \left(f\left(\frac{1}{g} - 2\sum_{i=1}^d \frac{\lambda_i}{f_i}\right) \right. \\ & \left. -1 - \log\left(f\left(\frac{1}{g} - 2\sum_{i=1}^d \frac{\lambda_i}{f_i}\right)\right)\right) \mathrm{d}\omega \\ \geq \;& 0\,. \end{aligned}$$

This proves that

$$I(\mathbf{y}) \leq \inf_{f\in\mathcal{M}_r} \left\{ K_\infty(f \mid g) \;:\quad y_i = \frac{1}{2\pi}\int_{\mathbb{T}} \frac{f}{f_i}\,\mathrm{d}\omega \text{ for all } i \right\}.$$

■

Lemma 3: [REPRESENTATION FORMULA FOR $\bar{I}$] For any $\boldsymbol{y} \in \mathbb{R}^d$, if $\bar{I}(\mathbf{y}) < \infty$, there exists $\boldsymbol{\lambda} \in \mathcal{D}_{\bar{\Lambda}}$ such that $\bar{I}(\mathbf{y}) = \langle \boldsymbol{\lambda}, \mathbf{y}\rangle - \bar{\boldsymbol{\Lambda}}(\boldsymbol{\lambda})$.

Corollary 2: For any $\boldsymbol{y} \in \mathbb{R}^d$, $\bar{I}(\mathbf{y}) < \infty$ if and only if $I(\mathbf{y}) < \infty$.

Proof: $\bar{\Lambda} \geq \Lambda$, so that $\bar{I} \leq I$. This proves that if $I(\mathbf{y}) < \infty$ then $\bar{I}(\mathbf{y}) < \infty$.
If now $I(\mathbf{y}) = +\infty$, there exists a sequence $\boldsymbol{\lambda}^m$ in $\mathcal{D}_\Lambda^\circ$ such that $\langle \boldsymbol{\lambda}^m, \mathbf{y}\rangle - \Lambda(\boldsymbol{\lambda}^m)$ tends to infinity. Since Λ is either continuous and finite on the boundary of $\mathcal{D}_\Lambda$, or tends to $-\infty$ on this boundary, $\| \boldsymbol{\lambda}^m \|$ tends to infinity. Now, let u be an accumulation point of $\boldsymbol{\lambda}^m/\|\boldsymbol{\lambda}^m\|$. For any m, any $\omega \in \mathbf{T}$,

$$\frac{1}{g(\omega)} - 2\sum_{i=1}^d \frac{\lambda_i^m}{f_i(\omega)} \geq 0$$

so that, since $\| \boldsymbol{\lambda}^m \|$ tends to infinity and all f_i are positive,

$$\sum_{i=1}^d \frac{u_i^m}{f_i(\omega)} \leq 0.$$

But by Lemma 1, this implies that $u \in \mathcal{D}_{\bar{\Lambda}}$, and also that for any positive M, $Mu \in \mathcal{D}_{\bar{\Lambda}}$. Now, as m tends to infinity,

$$\Lambda(\boldsymbol{\lambda}^m) \sim -1/2 \log \| \boldsymbol{\lambda}^m \|$$

so that $\langle \boldsymbol{\lambda}^m, \mathbf{y}\rangle \geq \mathbf{0}$ for large enough m, leading to $\langle u, \mathbf{y}\rangle \geq \mathbf{0}$. This implies that, for large enough M,

$$\langle Mu, \mathbf{y}\rangle - \bar{\Lambda}(Mu) \geq \frac{1}{3}\log M$$

so that $\bar{I}(\mathbf{y}) = +\infty$. ■

VII. TOOLS

When $\bar{I}(\cdot)$ and $I(\cdot)$ do not coincide, it is not possible to get a full analog of Lemma 2, that is to identify $\bar{I}()$ with an information divergence rate between an AR process and the AR process defined by g, nevertheless, thanks to Lemma 2 and 3, it is possible to get a partial information-theoretical interpretation of $\bar{I}(\cdot)$.

Lemma 4: Let $\mathbf{y}$ denote an element of $\mathcal{D}_I \subseteq \mathbb{R}^d$, let $\bar{\boldsymbol{\lambda}}$ be defined by $\bar{I}(\mathbf{y}) = \langle \bar{\boldsymbol{\lambda}}, \mathbf{y}\rangle - \bar{\Lambda}(\bar{\boldsymbol{\lambda}})$ and $\boldsymbol{\lambda}_\mathbf{y}$ be defined as the solution of $I(\mathbf{y}) = \langle \boldsymbol{\lambda}_\mathbf{y}, \mathbf{y}\rangle - \Lambda(\boldsymbol{\lambda}_\mathbf{y})$. For any $\boldsymbol{\lambda}$, let $f_{\boldsymbol{\lambda}}$ be defined by

$$\frac{1}{f_{\boldsymbol{\lambda}}} = \frac{1}{g} - 2\sum_{i=1}^d \frac{\lambda_i}{f_i}\,.$$

Then

$$\begin{aligned} & \bar{I}(\mathbf{y}) \\ = \;& I(\mathbf{y}) - K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\bar{\boldsymbol{\lambda}}}\right) \\ = \;& K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid g\right) - K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\bar{\boldsymbol{\lambda}}}\right) \\ = \;& K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid g\right) - \inf_{\boldsymbol{\lambda}\in\mathcal{D}_{\bar{\Lambda}}} K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\boldsymbol{\lambda}}\right)\,. \end{aligned}$$

The corrective term $\inf_{\boldsymbol{\lambda}\in\mathcal{D}_{\bar{\Lambda}}} K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\boldsymbol{\lambda}}\right)$ is the price to be paid for the lack of steepness of $\bar{\Lambda}$.

Proof: We first check that $\bar{I}(\mathbf{y}) = K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid g\right) - K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\bar{\boldsymbol{\lambda}}}\right)$.

$$\begin{aligned} & K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid g\right) - K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\bar{\boldsymbol{\lambda}}}\right) \\ = \;& \frac{1}{4\pi}\int_{\mathbb{T}} 2\sum_{i=1}^d f_{\boldsymbol{\lambda}_\mathbf{y}} \frac{\bar{\lambda}_i}{f_i} + \log\frac{g}{f_{\bar{\boldsymbol{\lambda}}}} d\omega \\ = \;& \sum_{i=1}^d \bar{\lambda}_i y_i + \frac{1}{4\pi}\int_{\mathbb{T}} \log\frac{g}{f_{\bar{\boldsymbol{\lambda}}}} d\omega \\ = \;& \langle \bar{\boldsymbol{\lambda}}, \mathbf{y}\rangle + \frac{1}{4\pi}\int_{\mathbb{T}} \log\left(1 - 2g\sum_{i=1}^d \frac{\bar{\lambda}_i}{f_i}\right) d\omega \\ = \;& \langle \bar{\boldsymbol{\lambda}}, \mathbf{y}\rangle - \bar{\Lambda}\left(\bar{\boldsymbol{\lambda}}\right) \\ = \;& \bar{I}(\mathbf{y})\,. \end{aligned}$$

Now, for any $\boldsymbol{\lambda} \in \mathcal{D}_{\bar{\Lambda}}$, $\langle \boldsymbol{\lambda}, \mathbf{y}\rangle - \Lambda\left(\bar{\boldsymbol{\lambda}}\right) \leq \bar{I}(\mathbf{y})$. But again, $\langle \boldsymbol{\lambda}, \mathbf{y}\rangle - \bar{\Lambda}(\boldsymbol{\lambda}) = K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid g\right) - K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\boldsymbol{\lambda}}\right)$, leading to $K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\boldsymbol{\lambda}}\right) \geq K_\infty\left(f_{\boldsymbol{\lambda}_\mathbf{y}} \mid f_{\bar{\boldsymbol{\lambda}}}\right)$. ■

Remark: If $d = 2$, $f_1 = g$ and $\lambda_2 = -\lambda_1$ with $\lambda_1 < 0$, the quadratic form under consideration is the stochastic part of the quasi-log-likelihood ratio between f_2 and g. The LDP satisfied by this log-likelihood ratio is well-understood thanks to Proposition 7 in [9] and Proposition 5.1 in [8]. It actually admits an information-theoretical representation property.

Lemma 5: Let $\mathbf{y}$ denote an element of $\mathcal{D}_I \subseteq \mathbb{R}^d$, let $\bar{\boldsymbol{\lambda}}$ be such that $\bar{I}(\mathbf{y}) = \left\langle \bar{\boldsymbol{\lambda}}, \mathbf{y}\right\rangle - \bar{\Lambda}(\bar{\boldsymbol{\lambda}})$, let $\bar{\mathbf{y}}$ be such that $I(\bar{\mathbf{y}}) = \left\langle \bar{\boldsymbol{\lambda}}, \bar{\mathbf{y}}\right\rangle - \Lambda(\bar{\boldsymbol{\lambda}})$, then

$$\bar{I}(\mathbf{y}) \geq I(\bar{\mathbf{y}})\,.$$

Proof: From the definitions of $\bar{\boldsymbol{\lambda}}$ and $\bar{\mathbf{y}}$, it follows that

$$\bar{I}(\mathbf{y}) - I(\bar{\mathbf{y}}) \geq \langle \bar{\boldsymbol{\lambda}}, \mathbf{y} - \bar{\mathbf{y}}\rangle\,.$$

For any $\boldsymbol{\lambda}$, we have

$$\langle \boldsymbol{\lambda}, \mathbf{y}\rangle - \bar{\Lambda}(\boldsymbol{\lambda}) \leq \left\langle \bar{\boldsymbol{\lambda}}, \mathbf{y}\right\rangle - \bar{\Lambda}(\bar{\boldsymbol{\lambda}}).$$

Now for any $\epsilon \in [0,1]$, $(1-\epsilon)\bar{\boldsymbol{\lambda}} \in \mathcal{D}_{\bar{\Lambda}}^\circ$, since both 0 and $\bar{\boldsymbol{\lambda}}$ belong to $\mathcal{D}_{\bar{\Lambda}}^\circ$. Substituting $(1-\epsilon)\bar{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$ in the preceding inequality, and rearranging leads to:

$$\frac{1}{\epsilon}\left(\bar{\Lambda}((1-\epsilon)\bar{\boldsymbol{\lambda}}) - \bar{\Lambda}(\bar{\boldsymbol{\lambda}})\right) \leq \left\langle \bar{\boldsymbol{\lambda}}, \mathbf{y}\right\rangle.$$

Letting ϵ tend to 0, and recalling that $\nabla\bar{\Lambda}(\boldsymbol{\lambda}) = \bar{\mathbf{y}}$:

$$\left\langle \bar{\boldsymbol{\lambda}}, \bar{\mathbf{y}} \right\rangle \leq \left\langle \bar{\boldsymbol{\lambda}}, \mathbf{y} \right\rangle \, .$$

■

The following lemma relates the shape of $\mathcal{D}_\Lambda$ and the shape of $\mathcal{D}_{\bar{\Lambda}}$ (clause a) of the Lemma), as a byproduct, we also get a relation between $\bar{I}(\mathbf{y})$ and a relative entropy with respect to g.

Lemma 6: Let ρ be defined as

$$\min_{\omega\in\mathbb{T}} g(\omega)/(2\max_{\omega\in\mathbb{T}} g\left(\omega\right)).$$

Let $(h_i)_{i=1,d}$ denote a sequence of functions from $\mathcal{H}$. Let I and $\bar{I}$ denote the rate functions associated with the sequence $(h_i)_{i=1,d}$. Let $\mathbf{y}$ denote and element of $\mathcal{D}_I \subseteq \mathbb{R}^d$. Let $\boldsymbol{\lambda} \in \mathcal{D}_\Lambda$ satisfy $I\left(\mathbf{y}\right) = \langle \boldsymbol{\lambda}, \mathbf{y} \rangle - \Lambda(\boldsymbol{\lambda})$. Let $f_{\boldsymbol{\lambda}}$ be defined by

$$\frac{1}{f_{\boldsymbol{\lambda}}} = \frac{1}{g} - 2\sum_{i=1}^{d} \lambda_i h_i \, .$$

Then the followings hold
a) $\rho\boldsymbol{\lambda} \in \mathcal{D}_{\bar{\Lambda}}$,
b) $\rho K_\infty\left(f_{\boldsymbol{\lambda}} \mid g\right) \leq \bar{I}\left(\mathbf{y}\right)$.

Proof: Let μ be such that $0 < \mu < 1$ and $\mu\boldsymbol{\lambda} \in \mathcal{D}_{\bar{\Lambda}}$. Note that

$$\frac{1}{f_{\mu\boldsymbol{\lambda}}} = \frac{\mu}{f_{\boldsymbol{\lambda}}} + \frac{1-\mu}{g} \, .$$

Then by the convexity of $K_\infty\left(f_{\boldsymbol{\lambda}} \mid \cdot\right)$:

$$K_\infty\left(f_{\boldsymbol{\lambda}} \mid f_{\mu\boldsymbol{\lambda}}\right) \quad \leq \quad (1-\mu) K_\infty\left(f_{\boldsymbol{\lambda}} \mid g\right) \, .$$

Now, let ρ be defined as in the statement of the Lemma. Then, for all tuples of functions h_i from $\mathcal{H}$, as $\boldsymbol{\lambda} \in \mathcal{D}_\Lambda$, for all $\omega \in \mathbb{T}$:

$$\frac{1}{\min_{\mathbb{T}} g(\omega)} \geq 2\sum_{i=1}^{d} \lambda_i h_i(\omega),$$

which entails $-2\rho\sum_{i=1}^{d} \lambda_i h_i \geq -1/\max_{\mathbb{T}} g(\omega)$. Hence, the largest eigenvalue of

$$T_n\left(2\rho\sum_{i=1}^{d} \lambda_i h_i\right)$$

is larger than $-1/\max_{\mathbb{T}} g(\omega)$, while the smallest eigenvalue of $T_n^{-1}\left(g\right)$ is not smaller than $1/\max_{\mathbb{T}} g(\omega)$. This finally entails that for all n

$$T_n^{-1}\left(g\right) - T_n\left(2\rho\sum_{i=1}^{d} \lambda_i h_i\right) = T_n\left(\frac{1}{f_{\rho\boldsymbol{\lambda}}}\right) + T_n^{-1}\left(g\right) - T_n\left(\frac{1}{g}\right)$$

is definite positive, which implies that $f_{\rho\boldsymbol{\lambda}} \in F$.

Then

$$\begin{aligned} \rho K_\infty\left(f_{\boldsymbol{\lambda}} \mid g\right) &\leq K_\infty\left(f_{\boldsymbol{\lambda}} \mid g\right) - K_\infty\left(f_{\boldsymbol{\lambda}} \mid f_{\rho\boldsymbol{\lambda}}\right) \\ &\leq \bar{I}\left(\mathbf{y}\right) \, . \end{aligned}$$

by Lemma 4. ■

The following partial information-theoretical interpretation of $\bar{I}$ is now an easy consequence of Lemma 2 and Lemma 4.

Lemma 7: Let $\mathbf{y}$ belong to $\mathcal{D}_{\bar{I}}$, let $f_{\mathbf{y}}$ be defined as the spectral density that minimizes $K_\infty\left(\cdot \mid g\right)$ among the solutions of

$$\frac{1}{2\pi}\int_{\mathbb{T}} \frac{f_{\mathbf{y}}}{f_i} = \mathbf{y}_i \, ,$$

then

$$\bar{I}\left(\mathbf{y}\right) \leq K_\infty\left(f_{\mathbf{y}} \mid g\right) - \inf_{h\in F} K_\infty\left(f_{\mathbf{y}} \mid h\right) \, .$$

Proof: From the representation formula for $I()$. ■

VIII. Under-estimation exponents

Throughout this section, g denotes the spectral density of the AR process of order $r > 0$ that generates the observations. Henceforth, $K_\infty\left(\mathcal{M}_{r-1} \mid g\right)$ denotes the minimum information divergence rate between AR-processes of order $r-1$ and the AR process with density g ($K_\infty\left(\mathcal{M}_{r-1} \mid g\right)$ is the Stein under-estimation exponent associated with g, see Theorem 5).

Sieve approximations allow to handle the large deviations of the log-likelihood processes indexed by $\mathcal{F}_p$ or equivalently by Θ^p using the LDPs for vectors of quadratic forms exhibited in Section V. Thanks to the sieve approximation Lemma (Lemma 8) and to the LDP upper-bound for vectors of quadratic forms 7, Lemma 3 provides a lower bound on order under-estimation exponent. This lower bound is defined as a limit of infima of large deviation rate functions. Such a definition does not preclude the triviality of the lower bound. The rest of this section is devoted to the identification of this limit with the expression given in the statement of Theorem 6 (Lemma 11) and to checking that the latter expression is non-trivial (Lemma 12).

Definition 11: [Sieve] For any integer p, any positive α, an α-sieve for the set $\mathcal{F}_p$ of spectral densities of stable AR(p)-processes with innovation variance equal to 1 is a finite subset $\mathcal{N}(\alpha)$ of $\mathcal{F}_p$, such that for any $f \in \mathcal{F}_p$, there exists $\hat{f} \in \mathcal{N}(\alpha)$ such that

$$\| \, \mathbf{a}_f - \mathbf{a}_{\hat{f}} \, \|_2 \leq \alpha,$$

where $\mathbf{a}_f$ (resp. $\mathbf{a}_{\hat{f}}$) is the prediction filter associated with f in Θ^p (resp. $\hat{f}$ in $\mathcal{F}_p$).

Upper bounds on the size of α-sieves for $\mathcal{F}_p$ or Θ^p can be checked by the following argument. If $\mathbf{a} \in \Theta^p$, then the complex polynomial $1 + \sum_{i=1}^{p} a_i z^i$ has no roots inside the open complex unit disc. Let us denote by $(z_i)_{i=1}^{p}$ its p complex roots,

$$1 + \sum_{i=1}^{d} a_i z^i = \prod_{i=1}^{p} (1 - z/z_i) \, .$$

This implies that for all $i \in \{1, \ldots, p\}$, $|a_i| \leq \binom{p}{i}$. Hence we get the following rough upper-bound on the minimal cardinality of an α-sieve for Θ^p:

$$\left(\frac{\sqrt{p}}{\alpha}\right)^p \prod_{i=1}^{p} \binom{p}{i} \leq \left(\frac{2^p}{\alpha}\right)^p \, .$$

Henceforth $N(\alpha)$ denotes the index set the sieve $\mathcal{N}(\alpha)$, by definition, we have:

$$\mathcal{N}(\alpha) = \{f_i \, ; \, f_i \in \mathcal{F}_r, \, i \in N(\alpha)\} \, .$$

In the sequel, $J(\cdot)$ is the rate function for the LDP satisfied by $\left(\frac{1}{n}\sum_{t=1}^{n} Y_t^2\right)_{n\in\mathbb{N}}$ when Y_t is generated by a Gaussian AR-(r) process with spectral density g, (we refer to [16] for a full presentation of this LDP and to the Appendix for an explicit definition of $J(\cdot)$).

Lemma 8: [SIEVE APPROXIMATION] For each $\alpha > 0$, let $\mathcal{N}(\alpha)$ denote an α-sieve for $\mathcal{F}_p$ according to Definition 11. $N(\alpha)$ denotes the index set for $\mathcal{N}(\alpha)$:

$$\mathcal{N}(\alpha) = \{f_i \ : \ f_i \in \mathcal{F}_p, i \in N(\alpha)\} \ .$$

For any positive M, any integer p, for any $\epsilon > 0$, if $\alpha > 0$ satisfies

$$\epsilon \geq \alpha(2r+1)2^{r+1}J^{-1}\left(M\right) \ ,$$

then

$$\limsup_{n\to+\infty} \frac{1}{n}\log P\left(\sup_{f\in\mathcal{F}_p}\inf_{i\in N(\alpha)}\left|\frac{1}{n}\mathbf{Y}^\dagger\left[T_n\left(\frac{1}{f}\right) - T_n\left(\frac{1}{f_i}\right)\right]\mathbf{Y}\right| \geq \epsilon\right) \leq -M$$

where $N(\alpha)$ denotes the index sets for α-sieves of $\mathcal{F}_p$.

The proof of the sieve approximation Lemma is given in the Appendix (Part *c*). The next definition is concerned with the sets of values for vectors of quadratic forms indexed by sieves for $\mathcal{F}_r$ that correspond to order under-estimation events.

Definition 12: For any $\alpha > 0$, let $N(\alpha)$ and $N'(\alpha)$ denote respectively the index sets of α-sieves of $\mathcal{F}_{r-1}$ and $\mathcal{F}_r$. The set C_α is the closed set of real vectors indexed by $N(\alpha)\times N'(\alpha)$ defined by:

$$C_\alpha = \left\{\mathbf{y} \ : \ \inf_{i\in N(\alpha)} y_i \leq \alpha(2r+1)2^{r+1}J^{-1}\left(K_\infty\left(\mathcal{M}_{r-1} \mid g\right)\right) + e^\alpha \times \inf_{j\in N'(\alpha)} y_j\right\}$$

and $\mathbf{y}^\alpha$ is an element of C_α that minimizes $\bar{I}$:

$$\mathbf{y}^\alpha \text{ such that } \bar{I}\left(\mathbf{y}^\alpha\right) = \inf_{\mathbf{y}\in C_\alpha} \bar{I}\left(\mathbf{y}\right) \ .$$

Notations $\bar{I}$ and I have been intentionally overloaded. For every index set $N(\alpha)\times N'(\alpha)$, they denote specific rate functions.

The existence of $\mathbf{y}^\alpha$ is ensured by the fact that $\bar{I}$ is lower-semi-continuous and C_α is closed.

Lemma 9: For each positive α, let $\mathbf{y}^\alpha$ be defined according to Definition 12, then

$$\limsup_{n\to+\infty}\frac{1}{n}\log\beta_n^{W,r}(P) \leq -\lim_{\alpha\searrow 0}\bar{I}\left(\mathbf{y}^\alpha\right) \ .$$

Proof: [Proof of Lemma 9]

$$\begin{aligned}
&\beta_n^{W,r}(P)\\
&\leq P\left(\exists p<r \ : \ \sup_{\sigma,f\in\mathcal{F}_p}\left\{\bar{\ell}_n(\sigma^2,f,\mathbf{Y}) - \text{pen}(n,p)\right\} \geq \sup_{\sigma,f\in\mathcal{F}_r}\left\{\bar{\ell}_n(\sigma^2,f,\mathbf{Y}) - \text{pen}(n,r)\right\}\right)\\
&= P\left(\exists p<r \ : \ \inf_{f\in\mathcal{F}_p}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\right\} \leq \inf_{f\in\mathcal{F}_r}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\right\} \times \exp\left[2\left(\text{pen}(n,p)-\text{pen}(n,r)\right)\right]\right) \ .
\end{aligned}$$

The second step follows from Proposition 1. Let ϵ be defined as $\alpha(2r+1)2^{r+1}J^{-1}\left(K_\infty\left(\mathcal{M}_{r-1}\mid g\right)\right)$. For any $\alpha > 0$, for large enough n, $\text{pen}(n,p) - \text{pen}(n,r) < \alpha$.

$$\begin{aligned}
&\beta_n^{W,r}(P)\\
&\leq P\left(\exists p<r \ : \ \inf_{f\in\mathcal{F}_p}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\right\} \leq \inf_{f\in\mathcal{F}_r}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\right\}\exp\alpha\right)\\
&= P\left(\inf_{f\in\mathcal{F}_{r-1}}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\right\} \leq \inf_{f\in\mathcal{F}_r}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f}\right)\mathbf{Y}\right\}\exp\alpha\right)\\
&\leq P\left(\inf_{i\in N(\alpha)}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f_i}\right)\mathbf{Y}\right\} \leq \epsilon + \inf_{i\in N'(\alpha)}\left\{\frac{1}{n}\mathbf{Y}^\dagger T_n\left(\frac{1}{f_i}\right)\mathbf{Y}\right\}\exp\alpha\right)\\
&\quad + P\left(\sup_{f\in\mathcal{F}_r}\inf_{i\in N(\alpha)}\left|\frac{1}{n}\mathbf{Y}^\dagger\left[T_n\left(\frac{1}{f}\right) - T_n\left(\frac{1}{f_i}\right)\right]\mathbf{Y}\right| \geq \epsilon\right).
\end{aligned}$$

The first summand on the right-hand-side is handled using the large deviations theorem (Theorem 7), while the second summand is handled using the sieve approximation lemma (Lemma 8). Altogether, this implies

$$\begin{aligned}
&\limsup_{n\to+\infty}\frac{1}{n}\log\beta_n^{ML,r}(P)\\
&\leq \max\left\{-\inf_{\mathbf{y}\in C_\alpha}\bar{I}\left(\mathbf{y}\right); \limsup_{n\to+\infty}\frac{1}{n}\log P\left(\sup_{f\in\mathcal{F}_r}\inf_{i\in N(\alpha)}\left|\frac{1}{n}\mathbf{Y}^\dagger\left[T_n^{-1}\left(f\right) - T_n^{-1}\left(f_i\right)\right]\mathbf{Y}\right| \geq \epsilon\right)\right\}\\
&\leq \max\left\{-\inf_{\mathbf{y}\in C_\alpha}\bar{I}\left(\mathbf{y}\right); -K_\infty\left(\mathcal{M}_{r-1}\mid g\right)\right\}.
\end{aligned}$$

But Theorem 5 (Stein upper-bound) imply that for small enough α,

$$\max\left\{-\inf_{\mathbf{y}\in C_\alpha}\bar{I}(\mathbf{y})\,;-K_\infty(\mathcal{M}_{r-1}\mid g)\right\}=-\inf_{\mathbf{y}\in C_\alpha}\bar{I}(\mathbf{y})\ ,$$

and the Lemma follows by letting α tend to 0. ∎

The next corollary follows from Theorem 5 (Stein upper-bound) and Lemma 9.

Corollary 3: Let $\mathbf{y}^\alpha$ be defined according to Definition 12, then

$$\lim_{\alpha\searrow 0}\bar{I}(\mathbf{y}^\alpha)\leq\inf_{f\in\mathcal{M}_{r-1}}K_\infty(f\mid g)\ .$$

Lemma 10: Let $\mathbf{y}^\alpha$ be defined according to Definition 12, let $\boldsymbol{\lambda}^\alpha\in\mathcal{D}_\Lambda$ be such that

$$I(\mathbf{y}^\alpha)=\langle\mathbf{y}^\alpha,\boldsymbol{\lambda}^\alpha\rangle-\Lambda(\boldsymbol{\lambda}^\alpha)\ .$$

Let σ^α be the innovation variance of $(f_{\boldsymbol{\lambda}^\alpha})_\alpha$ and $\mathbf{a}^\alpha$ its prediction filter in Θ^r. For any accumulation point $(\tilde{\sigma},\tilde{\mathbf{a}})\in(0,\infty)\times\Theta^r$, then $0<\tilde{\sigma}<\infty$, and if $\tilde{f}$ is the spectral density with innovation variance $\tilde{\sigma}$ and prediction filter $\tilde{\mathbf{a}}$, then $\tilde{f}$ is the spectral density of a stable AR (r) process.

Proof: [Proof of Lemma 10] Let $\rho=\frac{\min_{\mathbb{T}}g}{2\max_{\mathbb{T}}g}$, from Lemma 6, it follows that for any α,

$$\rho K_\infty(f_{\boldsymbol{\lambda}^\alpha}\mid g)\leq\bar{I}(\mathbf{y}^\alpha)\,.$$

But taking α to 0 and applying Corollary 3 leads to the fact that $\tilde{\sigma}$ cannot be 0 or $+\infty$, so that one obtains

$$K_\infty\left(\tilde{f}\mid g\right)\leq\frac{1}{\rho}\min_{f\in\mathcal{M}_{r-1}}K_\infty(f\mid g)\,.$$

This implies that $\tilde{f}$ defines a stable AR(r) process. ∎

Combining the next lemma with Lemma 3 proves Part b) of Theorem 6.

Lemma 11: Let $\mathbf{y}^\alpha$ be defined according to Definition 12, then

$$\lim_{\alpha\searrow 0}\bar{I}(\mathbf{y}^\alpha)\geq\inf_{f\in\mathcal{M}_{r-1}}\left[K_\infty(f\mid g)-\inf_{h\in F}K_\infty(f\mid h)\right]$$

.

Proof: [Proof of Lemma 11] According to Lemma 4, there exists $\boldsymbol{\lambda}^\alpha\in\mathcal{D}_\Lambda$ such that $f_{\boldsymbol{\lambda}^\alpha}$ is the spectral density of an AR-(r) process satisfying

$$\bar{I}(\mathbf{y}^\alpha)=K_\infty(f_{\boldsymbol{\lambda}^\alpha}\mid g)-\inf_{\boldsymbol{\lambda}\in\mathcal{D}_{\bar{\Lambda}}}K_\infty(f_{\boldsymbol{\lambda}^\alpha}\mid f_{\boldsymbol{\lambda}})$$

and

$$y_i^\alpha=\frac{1}{2\pi}\int_{\mathbb{T}}\frac{f_{\boldsymbol{\lambda}^\alpha}}{f_i}\text{ for }i\in I\cup J$$

which, since $\mathbf{y}^\alpha\in C_\alpha$ and all f_is verify $\int_{\mathbb{T}}\log f_i=0$, leads to

$$\min_{i\in I}\left(\frac{1}{2\pi}\int_{\mathbb{T}}\frac{f_{\boldsymbol{\lambda}^\alpha}}{f_i}\right)\leq e^\alpha\times\min_{i\in J}\left(\frac{1}{2\pi}\int_{\mathbb{T}}\frac{f_{\boldsymbol{\lambda}^\alpha}}{f_i}\right).\qquad(7)$$

Let $\tilde{\sigma}$, $\tilde{\mathbf{a}}$, $\tilde{f}$ be defined as in the statement of Lemma 10. There exists $m>0$ and M such that, for small enough α, we have $m\leq f_{\boldsymbol{\lambda}^\alpha}\leq M$ on $\mathbb{T}$, which, together with (7) leads to

$$\inf_{f\in\mathcal{F}_{r-1}}\left(\frac{1}{2\pi}\int_{\mathbb{T}}\frac{\tilde{f}}{f}\right)\leq\inf_{f\in\mathcal{F}_r}\left(\frac{1}{2\pi}\int_{\mathbb{T}}\frac{\tilde{f}}{f}\right)$$

and then to

$$\inf_{f\in\mathcal{M}_{r-1}}K_\infty\left(\tilde{f}\mid f\right)\leq\inf_{f\in\mathcal{M}_r}K_\infty\left(\tilde{f}\mid f\right).$$

This implies by Theorem 3 that $\tilde{f}$ is the spectral density of a stable AR $(r-1)$ process.

The Lemma will thus be proved as soon as the following is established:

$$\lim_{\alpha\searrow 0}\bar{I}(\mathbf{y}^\alpha)\geq K_\infty\left(\tilde{f}\mid g\right)-\inf_{h\in F}K_\infty\left(\tilde{f}\mid h\right).\qquad(8)$$

Define F_α as the set of spectral densities

$$f_{\boldsymbol{\lambda}}=\frac{g}{1-2\sum_{i=1}^{N(\alpha)}\lambda_i g/f_i}$$

with $\lambda\in\mathcal{D}_{\bar{\Lambda}}$ (the dimension of $\boldsymbol{\lambda}$'s depends on the size of the sieve $\mathcal{N}(\alpha)$, but for any α, F_α is a subset of F). Then for any $h\in F_\alpha$,

$$\bar{I}(\mathbf{y}^\alpha)\geq K_\infty(f_{\boldsymbol{\lambda}^\alpha}\mid g)-K_\infty(f_{\boldsymbol{\lambda}^\alpha}\mid h)\ .$$

Without loss of generality, we may assume that the sieves $\mathcal{N}(\alpha)$ are nested ($\alpha>\alpha'$ implies $\mathcal{N}(\alpha)\subseteq\mathcal{N}(\alpha')$). Hence for any α', for any $h\in F_{\alpha'}$

$$\lim_{\alpha\searrow 0}\bar{I}(\mathbf{y}^\alpha)\geq K_\infty\left(\tilde{f}\mid g\right)-K_\infty\left(\tilde{f}\mid h\right).$$

But for any h in F, with associated innovation variance σ_h^2 and prediction filter $\mathbf{a}_h$, there exists some $i\in N(\alpha)$ such that $h_\alpha\in F_\alpha$ has innovation variance σ_h^2 and prediction filter $\mathbf{a}_i$ such that $\parallel\mathbf{a}_h-\mathbf{a}_i\parallel\leq\alpha$, so that $\lim_{\alpha\to 0}K_\infty\left(\tilde{f}\mid h_\alpha\right)=K_\infty\left(\tilde{f}\mid h\right)$. Thus for any $h\in F$, one has

$$\lim_{\alpha\searrow 0}\bar{I}(\mathbf{y}^\alpha)\geq K_\infty\left(\tilde{f}\mid g\right)-K_\infty\left(\tilde{f}\mid h\right),$$

which leads to (8). ∎

Part a) of Theorem 6 follows from the next lemma.

Lemma 12: [NON TRIVIALITY OF UNDER-ESTIMATION EXPONENT]

$$\inf_{f\in\mathcal{M}_{r-1}}\left[K_\infty(f\mid g)-\inf_{h\in F}K_\infty(f\mid h)\right]>0.$$

Proof: [Proof of Lemma 12] Let h be the spectral density of a stable AR(r) process. For any real a such that $a > -\inf_{\mathbb{T}} h / \sup_{\mathbb{T}} g$, define h_a by

$$\frac{1}{h_a} = \frac{1}{g} + \frac{a}{h}.$$

Thus h_a is positive on $\mathbb{T}$ and is the spectral density of an AR(r) process. Moreover, the smallest eigenvalue of

$$T_n\left(\frac{1}{h_a}\right) + T_n^{-1}(g) - T_n\left(\frac{1}{g}\right) = a + T_n^{-1}(g)$$

is positive, so that $h_a \in F$. One has

$$\inf_{f \in \mathcal{M}_{r-1}} \left[K_\infty(f \mid g) - \inf_{h \in F} K_\infty(f \mid h) \right] \geq \inf_{f \in \mathcal{M}_{r-1}} \left[K_\infty(f \mid g) - K_\infty(f \mid h_a) \right],$$

and the infimum on the right-hand side is attained. Now, let $\tilde{f}$ be such that

$$\inf_{f \in \mathcal{M}_{r-1}} \left[K_\infty(f \mid g) - \inf_{h \in F} K_\infty(f \mid h) \right] \geq L(a) = K_\infty\left(\tilde{f} \mid g\right) - K_\infty\left(\tilde{f} \mid h_a\right).$$

But $L(a)$ satisfies the following equation:

$$L(a) = \frac{1}{4\pi} \int_{\mathbb{T}} \left[\log\left(1 + a\frac{g}{h}\right) - a\frac{\tilde{f}}{h} \right] \mathrm{d}\omega\,.$$

Hence $L()$ is a concave function, with $L(0) = 0$ and $L'(0) = \frac{1}{4\pi}\int_{\mathbb{T}} \frac{g-\tilde{f}}{h}\,\mathrm{d}\omega$. Now, $\tilde{f} \neq g$ since $\tilde{f}$ has order $\leq r-1$ and g has order r, and it is possible to choose h such that $L'(0) \neq 0$. Indeed, if this were not the case, we would get $\int_0^{2\pi} g(\omega)\cos(k\omega)\,\mathrm{d}\omega = \int_0^{2\pi} \tilde{f}(\omega)\cos(k\omega)\,\mathrm{d}\omega$ for $k = 0, \ldots, r$. But the spectral density of an AR(r) processes is determined by covariances with lags less than r. So, this would lead to $g = \tilde{f}$ and contradict the fact that $\tilde{f} \in \mathcal{F}_{r-1}$. Let h be such that $L'(0) \neq 0$. Then there exists a ($a < 0$ in case $L'(0) < 0$ and $a > 0$ in case $L'(0) > 0$) such that $L(a) > 0$. ∎

Acknowledgments. The authors would like to thank Jean Coursol and Thomas Duquesne for fruitful discussions.

REFERENCES

[1] R. Azencott and D. Dacunha-Castelle. *Séries d'observations irrégulières*. Masson, 1984.

[2] R.K. Bahr. Asymptotic analysis of error probabilities for the nonzero-mean Gaussian hypothesis testing problem. *IEEE Trans. Inform. Theory*, 36(3):597–607, 1990.

[3] Y. Baraud, F. Comte, and G. Viennet. Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.*, 5:33–49 (electronic), 2001.

[4] P. Barone, A. Gigli, and M. Piccioni. Optimal importance sampling for some quadratic forms of ARMA processes. *IEEE Trans. Inform. Theory*, 41(6, part 2):1834–1844, 1995.

[5] A. Barron. The strong ergodic theorem for densities; generalized Shannon-McMillan-Breiman theorem. *Annals of Probability*, 13:1292–1303, 1985.

[6] A. Barron, L. Birgé, and P. Massart. Risks bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113:301–415, 1999.

[7] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.

[8] B. Bercu, F. Gamboa, and M. Lavielle. Sharp large deviations for Gaussian quadratic forms with applications. *ESAIM Probab. Statist.*, 4:1–24 (electronic), 2000.

[9] B. Bercu, F. Gamboa, and A. Rouault. Large deviations for quadratic forms of stationary gaussian processes. *Stochastic Processes & their Applications*, 71:75–90, 1997.

[10] P.J. Bickel and K.J. Doksum. *Mathematical statistics*. Holden-Day Inc., San Francisco, Calif., 1976.

[11] A. Böttcher and B. Silbermann. *Introduction to large truncated Toeplitz matrices*. Universitext. Springer-Verlag, New York, 1999.

[12] M. Bouaziz. Inégalités de trace pour des matrices de Toeplitz et applications à des vraisemblances gaussiennes. *Probab. Math. Stat.*, 13(2):253–267, 1992.

[13] M. Bouaziz. Testing Gaussian sequences and asymptotic inversion of Toeplitz operators. *Probab. Math. Stat.*, 14(2):207–222, 1993.

[14] S. Boucheron and E. Gassiat. contributed chapter Order estimation in *Inference in Hidden Markov Models*. by O. Cappe, E. Moulines and T. Rydden. Springer-Verlag, 2005.

[15] P.J. Brockwell and R.A. Davis. *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer-Verlag, New York, 2002.

[16] W. Bryc and A. Dembo. Large deviations for quadratic functionals of Gaussian processes. *J. Theoret. Probab.*, 10(2):307–332, 1997.

[17] A. Chambaz. Testing the order of a model. *Ann. Statist.*, 2005. to appear.

[18] J. Coursol and D. Dacunha-Castelle. Sur la formule de Chernoff pour deux processus gaussiens stationnaires. *C. R. Acad. Sci., Paris, Ser. A*, pages 769–770, 1979.

[19] J. Coursol and D. Dacunha-Castelle. Remarks on the approximation of the likelihood function of a stationary Gaussian process. *Theory Probab. Appl.*, 27:162–167, 1982.

[20] T. Cover and J. Thomas. *Elements of information theory*. John Wiley & sons, 1991.

[21] I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44:2505–2523, 1998.

[22] I. Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48:1616–1628, 2002.

[23] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Academic Press, 1981.

[24] I. Csiszár and P. Shields. The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28:1601–1619, 2000.

[25] D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques*, volume 2. Masson, 1983.

[26] D. Dacunha-Castelle and E. Gassiat. The estimation of the order of a mixture model. *Bernoulli*, 3:279–299, 1997.

[27] D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes. *Annals of Statistics*, 27:1178–1209, 1999.

[28] A. Dembo and O. Zeitouni. Large deviations via parameter dependent change of measure, and an application to the lower tail of Gaussian processes. In *Seminar on Stochastic Analysis, Random Fields and Applications (Ascona, 1993)*, volume 36 of *Progr. Probab.*, pages 111–121. Birkhäuser, Basel, 1995.

[29] A. Dembo and O. Zeitouni. *Large deviation techniques and applications*. Springer, 1998.

[30] M. Feder and N. Merhav. Universal composite hypothesis testing: a competitive minimax and its applications. *IEEE Trans. Inform. Theory*, 48:1504–1517, 2002.

[31] L. Finesso, C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, 42:1488–1497, 1996.

[32] F. Gamboa, A. Rouault, and M. Zani. A functional large deviations principle for quadratic forms of gaussian stationary processes. *Statist. Prob. Letters*, 43:299–308, 1999.

[33] E. Gassiat and S. Boucheron. Optimal error exponents in hidden markov model order estimation. *IEEE Trans. Inform. Theory*, 49:864–880, 2003.

[34] V. Genon-Catalot and D. Picard. *Éléments de statistique asymptotique*, volume 11 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Paris, 1993.

[35] E. Hannan, J. McDougall, and D. Poskitt. Recursive estimation of autoregressions. *J. Roy. Statist. Soc. Ser. B*, 51(2):217–233, 1989.

[36] E. M. Hemerly and M. H. A. Davis. Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.*, 17(2):941–946, 1989.

[37] E. M. Hemerly and M. H. A. Davis. Recursive order estimation of autoregressions without bounding the model set. *J. Roy. Statist. Soc. Ser. B*, 53(1):201–210, 1991.

[38] T. Kailath and H.V. Poor. Detection of stochastic processes. *IEEE Trans. Inform. Theory*, 44(6):2230–2258, october 1998.

[39] Y. Kakizawa. On Bahadur asymptotic efficiency of the maximum likelihood and quasi-maximum likelihood estimators in Gaussian stationary processes. *Stochastic Processes & their Applications*, 85:29–44, 2000.

[40] S. Khudanpur and P. Narayan. Order estimation for a special class of hidden markov sources and binary renewal processes. *IEEE Trans. Inform. Theory*, 48:1704–1713, 2002.

[41] J.C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory*, 39:893–902, 1993.

[42] C. Léonard and J. Najim. An extension of Sanov's theorem: application to the Gibbs conditioning principle. *Bernoulli*, 8(6):721–743, 2002.

[43] C. Liu and P. Narayan. Order estimation and sequential universal data compression of a hidden markov source by the method of mixtures. *IEEE Trans. Inform. Theory*, 40:1167–1180, 1994.

[44] N. Merhav. The estimation of the model order in exponential families. *IEEE Trans. Inform. Theory*, 35(5):1109–1114, 1989.

[45] R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math.*, 35(3):415–423, 1983.

[46] M. Taniguchi and Y. Kakizawa. *Asymptotic theory of statistical inference for time series*. Springer Series in Statistics. Springer-Verlag, New York, 2000.

[47] A. van der Vaart and J.A. Wellner. *Weak convergence and Empirical Processes*. Springer-Verlag, 1996.

[48] A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

[49] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Trans. Inform. Theory*, 46:431–445, 2000.

[50] Y. Yang and A.R. Barron. An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory*, 44:95–116, 1998.

[51] J. Ziv and N. Merhav. Estimating the number of states of a finite-state source. *IEEE Trans. Inform. Theory*, 38:61–65, 1992.

APPENDIX

a) Stein exponents:

Proof: [Proof of triviality of over-estimation exponent] Let $g \in \mathcal{M}_{r-1}$ denote the spectral density of an AR-process of order $r-1$. Let P denote the distribution of the process.

Now let Q denote the probability distribution of an AR-process of order exactly r. The finite-dimensional projections Q^n of Q are absolutely continuous with respect to the finite-dimensional projections P^n of P. Let ϵ be positive and smaller than $\lim_n \beta_n^r(Q)$.. Let A_n^ϵ denote the event

$$A_n^\epsilon = \left\{ \mathbf{y} \ : \ \mathbf{y} \in \mathbb{R}^n, \ \frac{1}{n} \log \frac{dQ^n}{dP^n}(\mathbf{y}) \leq K_\infty \left(Q \mid P \right) + \epsilon \right\} .$$

By the Shannon-Breiman-McMillan Theorem (Theorem 1), for n large enough,

$$Q \left\{ A_n^\epsilon \right\} \geq 1 - \epsilon \, . \tag{9}$$

$$\begin{aligned}
\alpha_n^r(P) &= \mathsf{E}_P \left[\phi_n^r \right] \\
&= \mathsf{E}_Q \left[\frac{dP^n}{dQ^n} \phi_n^r \right] \\
&\geq \mathsf{E}_Q \left[\mathbf{1}_{A_n^\epsilon} \frac{dP^n}{dQ^n} \phi_n^r \right] \\
&\geq \exp \left(-n (K_\infty \left(Q \mid P \right) + \epsilon) \right) \left(\mathsf{E}_Q \left[\phi_n^r \right] - \epsilon \right) \\
&\geq \exp \left(-n (K_\infty \left(Q \mid P \right) + \epsilon) \right) \left(\beta_n^r(Q) - \epsilon \right) .
\end{aligned}$$

Taking the limit of logarithms with respect to n, as $\lim_n \beta_n^r(Q) > 0$:

$$\lim_n \frac{1}{n} \log \alpha_n^r(P) \geq -K_\infty \left(Q \mid P \right) - \epsilon \, . \tag{10}$$

Let σ^2 denote the variance of innovations associated to g ($\sigma^2 = 1/(2\pi) \int_{\mathbb{T}} g \, \mathrm{d}\omega$), and $\mathbf{b} \in \Theta^{r-1}$ the associated prediction filter. Let P denote the probability distribution associated with g.

Now, let $(\mathbf{a}^n)_{n \in \mathbb{N}}$ denote a sequence of elements of Θ^r such that for all n, $a_r^n \neq 0$

$$\sum_{i=1}^{r-1} (a_i^n - b_i)^2 \searrow 0 \quad \text{and} \quad a_r^n \searrow 0 \, .$$

Let P_n denote the probability distribution of the AR-process of order exactly r parameterized by $(\sigma, \mathbf{a}^n)$. Then $\lim_n K_\infty \left(P_n \mid P \right) = 0$. ∎

Proof: [Proof of Stein under-estimation exponent] Let P denote the probability distribution of an AR-process of order exactly r.

Now let Q denote the probability distribution of an AR-process of order $r-1$. The finite-dimensional projections Q^n of Q are absolutely continuous with respect to the finite-dimensional projections P^n of P. Let ϵ be positive and smaller than $\lim_n 1 - \alpha_n^r(Q)$. Let A_n^ϵ denote the event

$$A_n^\epsilon = \left\{ \mathbf{y} \ : \ \mathbf{y} \in \mathbb{R}^n, \ \frac{1}{n} \log \frac{dQ^n}{dP^n}(\mathbf{y}) \leq K_\infty \left(Q \mid P \right) + \epsilon \right\} .$$

By the Shannon-Breiman-McMillan Theorem (Theorem 1), for n large enough,

$$Q \left\{ A_n^\epsilon \right\} \geq 1 - \epsilon \, . \tag{11}$$

$$\begin{aligned}
\beta_n^r(P) &= \mathsf{E}_P \left[(1 - \phi_n^r) \right] \\
&\geq \mathsf{E}_P \left[\mathbf{1}_{A_n^\epsilon} (1 - \phi_n^r) \right] \\
&= \mathsf{E}_Q \left[\mathbf{1}_{A_n^\epsilon} \frac{dP^n}{dQ^n} (1 - \phi_n^r) \right] \\
&\geq \exp \left(-n (K_\infty \left(Q \mid P \right) + \epsilon) \right) \left(\mathsf{E}_Q \left[1 - \phi_n^r \right] - \epsilon \right) \\
&\geq \exp \left(-n (K_\infty \left(Q \mid P \right) + \epsilon) \right) \left(1 - \alpha_n^r(Q) - \epsilon \right) .
\end{aligned}$$

Taking the limit of logarithms with respect to n, as $\lim_n \alpha_n^r(Q) < 1$:

$$\lim_n \frac{1}{n} \log \beta_n^r(P) \geq -K_\infty \left(Q \mid P \right) - \epsilon \, . \tag{12}$$

Optimizing with respect to Q leads to the Theorem. ∎

Proof: [Scale invariance of error exponents] Let $\mathbf{a} \in \Theta^r \setminus \Theta^{r-1}$ denote a prediction filter of order exactly r. For all $\sigma > 0$, let P_σ denote the probability distribution of an AR process of order r parameterized by $(\sigma, \mathbf{a})$.

Note that the $\mathbb{R}^n$-valued random variable $\mathbf{Y}$ is distributed according to P_σ if and only if $1/\sigma \mathbf{Y}$ is distributed according to P_1 and that $f \in \mathcal{M}_r$ if and only if $f/\sigma^2 \in \mathcal{M}_r$.

The probability that the quasi-ML order testing procedure under-estimates the order on a random sample $\mathbf{Y}$ of length n is

$$\begin{aligned}
&P_\sigma\left\{e^{-\text{pen}(n,r)}\inf_{f\in\mathcal{M}_r}\mathbf{Y}^\dagger T_n^{-1}(f)\mathbf{Y}\right.\\
&\qquad\left.\leq e^{-\text{pen}(n,r-1)}\inf_{f\in\mathcal{M}_{r-1}}\mathbf{Y}^\dagger T_n^{-1}(f)\mathbf{Y}\right\}\\
&=\quad P_1\left\{e^{-\text{pen}(n,r)}\inf_{f\in\mathcal{M}_r}\frac{1}{\sigma}\mathbf{Y}^\dagger T_n^{-1}(f)\frac{1}{\sigma}\mathbf{Y}\right.\\
&\qquad\left.\leq e^{-\text{pen}(n,r-1)}\inf_{f\in\mathcal{M}_{r-1}}\frac{1}{\sigma}\mathbf{Y}^\dagger T_n^{-1}(f)\frac{1}{\sigma}\mathbf{Y}\right\}\\
&=\quad P_1\left\{e^{-\text{pen}(n,r)}\inf_{f\in\mathcal{M}_r}\mathbf{Y}^\dagger T_n^{-1}(\frac{f}{\sigma^2})\mathbf{Y}\right.\\
&\qquad\left.\leq e^{-\text{pen}(n,r-1)}\inf_{f\in\mathcal{M}_{r-1}}\mathbf{Y}^\dagger T_n^{-1}(\frac{f}{\sigma^2})\mathbf{Y}\right\}\\
&=\quad P_1\left\{e^{-\text{pen}(n,r)}\inf_{f\in\mathcal{M}_r}\mathbf{Y}^\dagger T_n^{-1}(f)\mathbf{Y}\right.\\
&\qquad\left.\leq e^{-\text{pen}(n,r-1)}\inf_{f\in\mathcal{M}_{r-1}}\mathbf{Y}^\dagger T_n^{-1}(f)\mathbf{Y}\right\}\ .
\end{aligned}$$

This is enough to conclude that

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\beta_n^{ML,r}(P_\sigma)\\
&\qquad\qquad=\limsup_{n\to\infty}\frac{1}{n}\log\beta_n^{ML,r}(P_1)\ .
\end{aligned}$$

The same line of reasoning works for the quasi-Whittle testing procedure and for the maximum likelihood testing procedure. ∎

b) Inverses of Toeplitz matrices associated with AR *processes:*
The next proposition provides with a quantitative assessment of the Whittle approximation when the symbol is the spectral density of an AR process.

Proposition 4: Let f denote the spectral density of an AR(r) process with unit innovation variance ($f\in\mathcal{M}_r$). Let $\mathbf{a}\in\Theta^r$ denote the associated prediction filter $f(z)=\frac{1}{\left|1+\sum_{i=1}^r a_i z^i\right|^2}$, let $a_0=1$. For $n\geq 2r$, the inverse of the Toeplitz matrix $T_n(f)$ is given by:

$$\begin{aligned}
&T_n^{-1}(f)[i,i+k]\\
&=\begin{cases}
0 & \text{if } |k|>r\\
\sum_{j=0}^{r-|k|}a_j a_{j+|k|} & \text{if } |k|\leq r \text{ and } i\wedge i+k>r\\
& \text{and } \ n-r\geq i\vee i+k\\
\sum_{j=0}^{(i\wedge i+k)-1}a_j a_{j+|k|} & \text{if } |k|\leq r \text{ and } i\wedge i+k\leq r\\
\sum_{j=0}^{n-(k+i\vee i)}a_j a_{j+|k|} & \text{if } |k|\leq r \text{ and } i\vee i+k>n-r
\end{cases}
\end{aligned}\tag{13}$$

The Toeplitz matrix associated with $1/f$ is given by:

$$T_n\left(1/f\right)[i,i+k]=\begin{cases}\sum_{j=0}^{r-|k|}a_j a_{j+|k|} & \text{if } |k|\leq r\\ 0 & \text{otherwise}\ .\end{cases}\tag{14}$$

Proof: Assume $(Y_i)_{i\in\mathbb{N}}$ is an AR(r) process with spectral density f.The log-likelihood of a vector $\mathbf{y}\in\mathbb{R}^n$ can be written in two different ways:

$$-\quad\frac{1}{2}\log\frac{1}{(2\pi)^n\text{det}(T_n(f))}\quad-\quad\frac{1}{2}\mathbf{y}^\dagger T_n^{-1}(f)\mathbf{y}$$

and

$$\begin{aligned}
&-\frac{1}{2}\log\frac{1}{(2\pi)^n\text{det}(T_n(f))}-\frac{1}{2}\mathbf{y}_{1:r}^\dagger T_r^{-1}(f)\mathbf{y}_{1:r}\\
&\qquad\qquad-\frac{1}{2}\sum_{t=r+1}^{n}\left(y_t+\sum_{i=1}^{r}a_i y_{t-i}\right)^2\ .
\end{aligned}$$

Note that $T_n^{-1}(f)$ is symmetric with respect to its two diagonals. Identifying coefficients of $y_t y_s$in the two preceding expressions leads to Equation (13).

Equation (14) follows immediately from the definition of T_n and $1/f$. ∎

Proposition 5: Let f denote the spectral density of an AR-(r) process with unit innovation variance, let $\mathbf{a}$ denote the associated prediction filter, let $a_0=1$, then:
a) $T_n\left(1/f\right)-T_n^{-1}(f)$ is non-negative of rank at most $2r$.
b)

$$\mathbf{y}^\dagger\left[T_n\left(1/f\right)-T_n^{-1}(f)\right]\mathbf{y}\leq r\left\|\mathbf{a}\right\|_2^2\left(\sum_{t=1}^{r}y_t^2+\sum_{t=n-r+1}^{n}y_t^2\right).\tag{15}$$

Proof: The following equation follows from Proposition 4:

$$\begin{aligned}
&\mathbf{y}^\dagger\left[T_n\left(\frac{1}{f}\right)-T_n^{-1}(f)\right]\mathbf{y}\\
&\quad=\sum_{j=1}^{r}\left\{\left(\sum_{i=j}^{r}a_i y_{n+j-i}\right)^2+\left(\sum_{i=j}^{r}a_i y_{i+1-j}\right)^2\right\}
\end{aligned}\tag{16}$$

Then Proposition 5 follows from Cauchy-Schwarz inequality. ∎

c) Proof of the sieve approximation lemma 8 :
Proof: [Sieve approximation] As both $\left(T_n^{-1}(f)-T_n^{-1}(f_i)\right)$ and $T_n\left(\frac{1}{f}-\frac{1}{f_i}\right)$ are band-limited matrices, let us first get a general upper-bound on

$$\mathbf{y}^\dagger A\mathbf{y}$$

where A is a $n\times n$ symmetric matrix such $A[i;i+k]=0$ whenever $|k|>r$. Agreeing on the fact that $A[i;j]=0$

whenever either i or j does not belong to $\{1,\dots,n\}$, we get:

$$\begin{aligned}
&\mathbf{y}^\dagger A\mathbf{y}\\
&= \sum_{i=1}^{n}\sum_{k=-r}^{r} y_i y_{i+k} A[i;i+k]\\
&= \sum_{i=1}^{n}\left(y_i \sum_{k=-r}^{r} y_{i+k} A[i;i+k]\right)\\
&\le \left(\sum_{i=1}^{n} y_i^2\right)^{1/2}\left(\sum_{i=1}^{n}\left(\sum_{k=-r}^{r} y_{i+k} A[i;i+k]\right)^2\right)^{1/2}\\
&\le \|\mathbf{y}\|\left(\sum_{i=1}^{n}\left(\sum_{k=-r}^{r} A^2[i;i+k]\right)\left(\sum_{k=-r}^{r} y_{i+k}^2\right)\right)^{1/2}\\
&\le \|\mathbf{y}\|\ \max_i\left(\sum_{k=-r}^{r} A^2[i;i+k]\right)^{1/2}\left(\sum_{i=1}^{n}\left(\sum_{k=-r}^{r} y_{i+k}^2\right)\right)^{1/2}\\
&\le \|\mathbf{y}\|\ \max_i\left(\sum_{k=-r}^{r} A^2[i;i+k]\right)^{1/2}\sqrt{2r+1}\left(\sum_{i=1}^{n} y_i^2\right)^{1/2}\\
&\le \max_{i\in\{1,\dots,n\}}\left(\sum_{k=-r}^{r} A^2[i;i+k]\right)^{1/2}\sqrt{2r+1}\|\mathbf{y}\|_2\ .
\end{aligned}$$

Now from Proposition 4, it follows that if $\mathbf{a}$ and $\mathbf{b}$ denote the prediction filters associated with f and f_i from $\mathcal{F}_r$:

$$\begin{aligned}
&\left(T_n^{-1}(f)[l;l+k]-T_n^{-1}(f_i)[l;l+k]\right)^2\\
&\le \left(\sum_{l=0}^{r-|k|} a_l a_{l+k}-b_l b_{l+k}\right)^2\\
&\le 2\left(\left(\sum_{l=0}^{r-|k|} a_l(a_{l+k}-b_{l+k})\right)^2\right.\\
&\qquad \left.+\left(\sum_{l=0}^{r-|k|} b_{l+k}(a_l-b_l)\right)^2\right)\\
&\le 2\left(\sum_{l=0}^{r}(a_l-b_l)^2\right)\ \left(\|\mathbf{a}\|^2\vee\|\mathbf{b}\|^2\right)\\
&\le 2^{2r+1}\alpha^2\,.
\end{aligned}$$

Plugging this bound in the preceding inequality, we get

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_r}\inf_{i\in N(\alpha)}\frac{1}{n}\left|\mathbf{Y}^\dagger\left(T_n^{-1}(f)-T_n^{-1}(f_i)\right)\mathbf{Y}\right|\\
&\qquad\le \frac{1}{n}\alpha(2r+1)2^{r+1}\left\|\mathbf{Y}\right\|^2\ .
\end{aligned}$$

In a similar way, we also get:

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_r}\inf_{i\in N(\alpha)}\frac{1}{n}\left|\mathbf{Y}^\dagger T_n\left(\frac{1}{f}-\frac{1}{f_i}\right)\mathbf{Y}\right|\\
&\qquad\le \frac{1}{n}\alpha(2r+1)2^{r+1}\left\|\mathbf{Y}\right\|^2\ .
\end{aligned}$$

Hence it remains to check that for any positive M, there exists an integer t such that

$$\limsup_{n\to\infty}\frac{1}{n}\log P\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i^2\ge\frac{1}{t}\right\}\le -M$$

From Theorem 2 in [16], it follows that, $\frac{1}{n}\sum_{t=1}^{n} Y_t^2$ satisfies a LDP with rate function

$$J(y)=\sup_{\lambda\in(-\infty,1/2\|g\|_\infty]}\{\lambda y-L(\lambda)\}$$

where

$$L(\lambda)=-\frac{1}{4\pi}\int_{\mathbb{T}}\log\det\left(1-2\lambda g(\omega)\right)\,\mathrm{d}\omega\,.$$

Now in order to terminate the proof of Lemma 8, it is enough to choose α in such a way that

$$J\left(\frac{\epsilon 2^{-(r+1)}}{\alpha(2r+1)}\right)\ge M\,.$$

■

d) Non-triviality of $F(g)$ for AR-processes of order 2:

Proof: [Proposition 3] Let g denote the spectral density of an auto-regressive process of order 2 with prediction filter (a_1,a_2) then

$$f(g)\in F\Leftrightarrow(1+a_2)^2>a_1^2(1+a_2^2).$$

Let $g\in\mathcal{F}_2$ be defined by the prediction filter $\mathbf{a}=(a_1,a_2)$. Let us agree on the following notations

$$\begin{aligned}
b_0 &= \sqrt{1+a_2^2}\\
b_1 &= \sqrt{1+a_2^2}\frac{a_1}{1+a_2}\\
P_a(z) &= 1+a_1 z+a_2 z^2.
\end{aligned}$$

The I-projection $f(g)$ of g on $\mathcal{M}_1$ is the spectral density $1/|b_0+b_1 z|^2$.

If $|z|\leqslant 1$, then $P_a(z)\neq 0$. Hence $P_a(x)$ is non-negative on $[-1,1]$. Moreover $P_a(1)>0$ and $P_a(-1)>0$. This implies

$$-(1+a_2)<a_1<1+a_2$$

which entails $|b_1|<b_0$.

The matrix $T_n(1/f(g))+T_n^{-1}(g)-T_n(1/g)$ coincides with

$$M_n=\begin{pmatrix}
d_1 & e & 0 & & & & & & 0\\
e & d_2 & c & 0 & & & & & \\
0 & c & d & c & & & & & \\
 & 0 & c & \ddots & \ddots & & & & \\
 & & 0 & \ddots & & \ddots & \ddots & \ddots & \\
 & & & & & & & \ddots & \\
 & & & & & \ddots & d & c & 0\\
 & & & & & \ddots & c & d_2 & e\\
0 & & & & & & 0 & e & d_1
\end{pmatrix}$$

where

$$\left\{\begin{array}{l} d = b_0^2 + b_1^2 \\ d_2 = b_0^2 + b_1^2 - a_2^2 \\ d_1 = b_0^2 + b_1^2 - a_1^2 - a_2^2 \\ c = b_0 b_1 \\ e = b_0 b_1 - a_1 a_2. \end{array}\right.$$

M_n is definite positive if all determinants of its "main minors" are positive. Let E_k denote the determinant of the sub-matrix formed by the last k rows and columns of M_n.

1) If $k \leqslant n-2$ and $k \geqslant 3$. Then $E_k = dE_{k-1} - c^2 E_{k-2}$. The polynomial $z^2 - dz + c^2$ has roots $\frac{d \pm (b_0^2 - b_1^2)}{2}$. This entails

$$E_k = \alpha b_0^{2k} + \beta b_1^{2k}$$

with

$$\left\{\begin{array}{l} \alpha b_0^2 + \beta b_1^2 = d_1 \\ \alpha b_0^4 + \beta b_1^4 = d_1 d_2 - e^2 \end{array}\right.$$

Note that

$$d_1 = \frac{(1+a_2)^2 - 2a_2 a_1^2}{(1+a_2)^2}$$

and that the condition $d_1 > 0$ may not be satisfied, for example if $a_2 = r^2$, $a_1 = -2r$ and $r \sim 1 - \epsilon$. Note that $e = a_1(1-a_2)/(1+a_2)$ and $d_2 = 1 + b_1^2$.
In order to have $E_k \geqslant 0$ for all k, it is necessary that $\alpha > 0$, hence $d_1 d_2 - e^2 - d_1 b_1^2 > 0$ and $d_2 - b_1^2 = 1$, that is $d_1 - e^2 > 0$.
This finally entails the necessary condition

$$(1+a_2)^2 > a_1^2(1+a_2^2).$$

2) The case $k = n-1$, boils down to the following relation

$$\begin{array}{rcl} E_{n-1} & = & d_2 E_{n-2} - c^2 E_{n-3} \\ & = & (d_2 b_0^2 - c^2)\alpha b_0^{2(n-3)} + (d_2 b_1^2 - c^2)\beta b_1^{2(n-3)} \\ & = & \alpha b_0^{2(n-3)} + (1 - b_1^2(b_0^2 - b_1^2)\beta b_1^{2(n-3)}) \\ & > & 0. \end{array}$$

3) Finally the case $k = n$ is dealt with by

$$\begin{array}{rcl} E_n & = & d_1 E_{n-1} - e^2 E_{n-2} \\ & = & (d_1 d_2 - e^2) E_{n-2} - d_1 c^2 E_{n-3} \\ & = & \left[(d_1 d_2 - e^2) b_0^2 - d_1 c^2\right] \alpha b_0^{2(n-3)} \\ & & + \left[(d_1 d_2 - e^2) b_1^2 - d_1 c^2\right] \beta b_1^{2(n-3)} \\ & > & 0 \\ & & \text{for sufficiently large } n \\ & & \text{as soon as } (d_1 d_2 - e^2) b_0^2 - d_1 c^2 > 0 \end{array}$$

But

$$\begin{array}{l} (d_1 d_2 - e^2) b_0^2 - d_1 c^2 \\ \quad = \quad b_0^2 \left[d_1 d_2 - e^2 - d_1 b_1^2\right] \\ \quad = \quad b_0^2 \left[d_1 (d_2 - b_1^2) - e^2\right] \\ \quad = \quad b_0^2 (d_1^2 - e^2) \quad \text{since } d_2 = 1 + b_1^2. \end{array}$$

Hence, $f(g) \in F(g)$ if and only if $(1+a_2)^2 > a_1^2(1+a_2^2)$. ■

Elisabeth Gassiat received the Doctorat in Mathematics degree in 1988 from Université Paris XI, France. She served as a Professor at Université Evry-Val d'Essone from 1993 till 1998. She is currently Professor of Mathematics at Université Paris XI, France. Her areas of interest include semi-parametric statistics, concentration inequalities, mixture and hidden Markov modeling and deconvolution.

Stéphane Boucheron received the Doctorat in Computer Science degree in 1988 from the Université de Montpellier (II), France.

During 1989-1990, he worked at Alcaltel-Alstom Corporate Research Center in Marcoussis, France. From 1991 till 2004, he has been with CNRS at Laboratoire de Recherche en Informatique, Université Paris XI, France. He is currently Professor of Mathematics at Université Paris VII-Denis Diderot, France. His main fields of interest are communication theory, statistical learning and random structures and algorithms.