The Annals of Statistics
1996, Vol. 24, No. 5, 1964–1981

BLIND DECONVOLUTION OF DISCRETE LINEAR SYSTEMS

BY F. GAMBOA AND E. GASSIAT

Laboratoire de Statistiques Orsay and Université Paris-Nord, and Université d'Evry Val d'Essonne

We study the blind deconvolution problem in the case where the input noise has a finite discrete support and the transfer linear system is not necessarily minimum phase. We propose a new family of estimators built using algebraic considerations. The estimates are consistent under very wide assumptions: The input signal need not be independently distributed; the cardinality of the finite support may be estimated simultaneously. We consider in particular AR systems: In this case, we prove that the estimator of the parameters is perfect a.s. with a finite number of observations.

1. Introduction.

Here we consider the linear process $Y := (Y_t)_{t \in Z}$,

$$Y_t = \sum_{k \in \mathbb{Z}} u_k X_{t-k}, \tag{1.1}$$

where $u = (u_k)_{k \in Z}$ is a deterministic filter, Z and $X = (X_t)_{t \in Z}$ are random series and $\mathbb{Z}$ is the set of signed integers. In this system, the variables X_t are unobservable, the filter u is unknown and the observations are the Y_t, $t = 1, \dots, n$.

We will make the following assumptions throughout the paper:

(M1) X_t, $t \in \mathbb{Z}$, are discrete, real with unknown common support $A := \{x_1, \dots, x_p\}$, where $x_1 < x_2 < \cdots < x_p$ and $p \geq 2$.

(M2) Let $U(x) := \sum_{k \in \mathbb{Z}} u_k e^{ikx}$. U is a continuous function and does not vanish on $[0, 2\pi)$. $\theta = (\theta_k)_{k \in \mathbb{Z}}$ is the inverse filter of u; that is,

$$\sum_{j \in \mathbb{Z}} \theta_j u_{k-j} = \delta_k, \qquad k \in \mathbb{Z},$$

where δ_k denotes the Kronecker symbol.

The problem of the estimation of the unknown inverse filter θ (and simultaneously the restitution of the input X_t by inversion) using only the observations Y_t is known as the blind deconvolution problem.

In this paper, we do not assume that the system is causal; that is, u_k may be nonzero for negative integers k. When the sequence $(X_t)_{t \in Z}$ is independent identically distributed (i.i.d.), and with a causal system, classical linear

Received July 1993; revised September 1995.
AMS 1991 *subject classifications*. 62G05, 62M09, 62M10.
Key words and phrases. Deconvolution, contrast function, T-system, discrete linear systems.

prediction can be used [see Azencott and Dacunha-Castelle (1984) for details and further references]. Linear prediction is optimal in the Gaussian case. Otherwise, Kreiss proposed adaptive estimators that are asymptotically minimax for causal ARMA processes [see Kreiss (1987)]. When the system is not a priori causal, other procedures of estimation exist. In general, these procedures use higher order moments [for example, cumulants; see Gassiat (1990)] of the observations or empirical spectral functions [not only spectral density, see Lii and Rosenblatt (1982)]. Indeed, second order moments are unable to distinguish between systems with the same spectral density but with different phases. Optimal procedures are investigated in Gassiat (1993) for regular models (i.e., for models where the distribution of X_1 is absolutely continuous with respect to Lebesgue measure). Gassiat (1993) proves asymptotic minimax lower bounds in the general noncausal case and shows how the causality allows adaptivity or not.

In this paper, we study the case where X is a discrete-valued process. This is a case of interest especially in digital communications [see Feher (1987)], where X stands for the transmitted signal, u represents the communication channel and Y is the signal observed at the receiver (cellular telephone, high-definition satellite, ...). For such situations, and if the process X is i.i.d., earlier nonminimum phase estimations may be used, as described in Gassiat (1990). However, these procedures, since they are adapted to nearly any distribution of X_t, do not incorporate the information of discreteness of the input signal and are consequently far from being optimal. A deconvolution method has recently been proposed in Li (1992, 1995) for multilevel inputs. This method presents the advantage of being able to handle nonstationary independent inputs (that is, independent but not necessarily identically distributed inputs). However, this method involves knowledge of the set A in which the values of X_t lie. In some sense, it is not really a *blind* method.

We propose in this work a new and powerful method of estimation that incorporates only the discreteness as prior information:

1. The estimators are proved in Theorem 2.8 to be asymptotically consistent under very wide assumptions: *The input signal does not need to be independently distributed*; *the support A does not need to be known*.
2. The method can be adapted to estimate simultaneously the cardinality of A without any upper bound on it, and leads to consistent estimators; see Theorem 4.2.
3. For particular systems such as AR systems, the estimator is a.s. perfect with a finite number of observations (see Theorem 3.1) even when the input signal is neither stationary nor independently distributed.
4. Though the method is very general, examples of estimators are very simple, using, for example, Hankel forms or Toeplitz forms. In these cases, the method requires the computation of the determinant of a matrix whose dimension is related to the cardinality of A, which is not very big in most applications (2–32 symbols depending on the alphabet, that is, the set A, which is used in the transmission).

5. The method may be extended to deal with noisy observations, where some unknown noise is added to the observation Y_t. This point is developed in Gassiat and Gautherat (1994), where numerical simulations show the efficiency of the method in various situations.
6. Similar ideas apply in different contexts: for instance, to estimate the number of components in a mixture, see Dacunha-Castelle and Gassiat (1994) or, for the problem of source separation when the source is discrete, see Gamboa and Gassiat (1995).
7. Our work clarifies the structure of linear discrete models, that is, why and how this structure can be exploited to obtain estimators converging very fast.

Our work originates in two very simple remarks:

REMARK 1.1. When adding two discrete variables, the support of the resulting variable is larger than the original ones.

REMARK 1.2. We are able to propose functions that distinguish between discrete variables with support of cardinality p and the others (they will be described further).

Now, if we consider for a possible inverse filter s the filtered series $Z(s) = (Z_t(s))_{t \in \mathbb{Z}}$,

$$Z_t(s) := \sum_{k \in \mathbb{Z}} s_k Y_{t-k} = \sum_{k \in \mathbb{Z}} (s * u)_k X_{t-k}, \qquad t \in \mathbb{Z}, \tag{1.2}$$

then θ is the only value of s such that $Z_t(s)$ is a discrete variable with support of cardinality p (see Theorem 2.2). θ may then be recovered through the investigation of the support of $Z(s)$, which is systematically done using the previously announced functions of the variables $Z_t(s)$.

The paper is organized as follows: in Section 2, we give the assumptions and the construction of the estimator $\hat{\theta}$ of θ. The main theorem is then Theorem 2.8 that states the convergence of the estimator. In Section 3, we do not assume that X is stationary. We state precisely the speed of convergence for the special case of autoregressive processes: Perfect estimation is achieved with a finite number of observations, which is specified. In Section 4, we explain how it is possible to estimate simultaneously the inverse filter θ and the cardinality of A. In Section 5, we explore how our results may be extended to other models. All proofs are collected in Section 6.

2. The estimation method. In this section, we explain how the structure of the model allows a very simple characterization of the inverse filter. For a while, we assume that the cardinality p of the support A is known.

2.1. *Assumptions*. We first give our assumptions on the process X:

(M3) X is a stationary ergodic process.

(M4) For any integer n and for any integers $j_1, \dots, j_n$ in $\{1, \dots, p\}$,

$$P(X_1 = x_{j_1}, \dots, X_n = x_{j_n}) > 0.$$

Assumption (M3) allows us to approximate all expectations of linear processes constructed with X through the empirical means. Assumption (M4) is sufficient to formalize Remark 1.1 in the following way:

PROPOSITION 2.1. *Assume that* (M1) *and* (M4) *hold. Let $a \in l^1(\mathbb{Z})$ be a filter with at least two nonzero coefficients. Let $W(a) := \sum_{k \in \mathbb{Z}} a_k X_k$. Then there exist $p + 1$ disconnected intervals $I_1, \dots, I_{p+1}$ such that*

$$P(W(a) \in I_j) > 0, \qquad j = 1, \dots, p+1.$$

Let us give simple examples where the assumptions hold:

1. *White noise*: When the variables X_t, $t \in \mathbb{Z}$, are independent identically distributed, (M1), (M3) and (M4) hold.
2. *Thresholded process*: Let $(W_t)_{t \in \mathbb{Z}}$ be a stationary process such that for any $j > 0$ and $t_1 \leq t_2 \leq \cdots \leq t_j$, the distribution of $(W_{t_i})_{i=1,\dots,j}$ has a positive density with respect to the Lebesgue measure on $\mathbb{R}^j$ (for example, a Gaussian process). Let $m_1 < m_2 < \cdots < m_{p-1}$ be real numbers and set $m_0 = -\infty$, $m_p = +\infty$. Define the thresholded process $(X_t)_{t \in \mathbb{Z}}$ by $X_t = x_j$ if and only if $W_t \in (m_j, m_{j+1}]$. Then (M1), (M3) and (M4) hold.
3. *Aperiodic recurrent Markov chain*: Let $(X_t)_{t \in \mathbb{Z}}$ be a Markov chain with state space of cardinality p and a transition matrix such that any transition probability is positive (all states communicate in one step). Then (M1) and (M4) hold. It is also easy to see that the chain is aperiodic and recurrent, so that (M3) holds.

Proposition 2.1 gives the following simple characterization of the inverse filter θ:

THEOREM 2.2. *For any summable filter s, the random variable $Z_1(s)$ is discrete with at most p points of support if and only if s is the inverse θ up to scale and delay*, *that is*,

$$\exists K \in \mathbb{Z}, \exists \lambda \in \mathbb{R}, \forall k \in \mathbb{Z}, \qquad s_k = \lambda \cdot \theta_{k-K}.$$

Theorem 2.2 is an obvious consequence of Proposition 2.1 and (1.2).

The deconvolution problem is ambiguous on the scale and the delay of the filter. Indeed, if

$$\exists r > 0, \exists K \in \mathbb{Z}, \forall k \in \mathbb{Z}, \qquad v_k = r u_{k-K}, \qquad X'_k = \frac{1}{r} X_{k-K},$$

then the series $Y'_t := \sum_k v_k X'_{t-k}$ satisfies $Y' = Y$, so that we have to fix the scale and the delay. Hence, we define the parameter space Θ as a subset of $l_1(\mathbb{Z})$ which is unambiguous on scale and delay. Here, unambiguous means that Θ is a subset of $l_1(\mathbb{Z})$ that does not contain two different filters s, s' such that

$$\exists r \in \mathbb{R}^*, \exists K \in \mathbb{Z}, \forall k \in \mathbb{Z}, \qquad s_k = r s'_{k-K}.$$

For example, we may take for Θ a subset of

$$\{s = (s_k)_{k \in \mathbb{Z}} \in l_1(\mathbb{Z}): s_0 = 1, |s_k| < 1, k > 0, |s_k| \le 1, k < 0\}.$$

The method we propose to estimate the inverse filter θ is derived from contrast estimation. To do this, we need to define a contrast function $H(s)$, defined for filters s in the space Θ of filters, and that achieves its unique minimum at the true value θ of the inverse filter. We also need a sequence $H_n(s)$ of random functions that involves only the observations and that converges (uniformly enough) to the contrast function H. Then, if the estimator $\hat{\theta}$ is defined as any minimizer of $H_n(s)$ on a compact space of filters, this estimator converges asymptotically to the true value of θ.

2.2. *The contrast function.*

2.2.1. *Definition of the contrast function.* $H(s)$ has only to discriminate between discrete variables with support of cardinality p and the others. In other words, $H(s)$ has to be a systematic computational description of the support of the variable $Z_1(s)$. To do this, let $\Phi = (1, \Phi_1, \dots, \Phi_{2p})$ be a Tchebychev system (T-system) of functions on $[0, 1]$ [for the definition of T-system, see Krein and Nudel' man (1977)]. A classical example consists of the p first sine and cosine functions. For any filter s, define

$$c(s) = (c^i(s))_{i=1,\dots,2p},$$
$$c^i(s) = E[\Phi_i(\varphi(Z_1(s)))],$$

where φ is a given continuous bijective function which maps $\mathbb{R}$ onto $(0, 1)$. Notice that $\varphi(Z_1(s))$ is now a variable taking value in $(0, 1)$. Now, a nice property of a T-system is the following: Let $\mathscr{P}$ be the set of all probability measures on $[0, 1]$ and

$$\mathscr{K} := \left\{c \in \mathbb{R}^{2p}: \exists P \in \mathscr{P}, \int_0^1 \Phi \, dP = c\right\}.$$

THEOREM 2.3. *If V is a random variable taking value in* $(0, 1)$, *then* $E(\Phi(V))$ *lies on the boundary of $\mathscr{K}$ if and only if V is discrete with at most p points of support.*

The proof is immediate by applying Theorem 4.1 of Krein and Nudel'man [(1977), page 78].

Let now h be a nonnegative and continuous function defined on $\bar{\mathscr{K}}$ such that

$$h(c) = 0 \quad \Leftrightarrow \quad c \in bd(\mathscr{K}).$$

Examples of such functions follow. We then define our function H.

DEFINITION 2.4.

$$H(s) = h(c(s)), \qquad s \in \Theta.$$

The following result states that H is a contrast function:

THEOREM 2.5. *Assume that* (M1), (M2) *and* (M4) *hold. For any filter s in the parameter space Θ, $H(s)$ is nonnegative. Moreover*,

$$H(s) = 0 \quad \Leftrightarrow \quad s = \theta.$$

2.3. *Examples of contrast functions.*

2.3.1. *Toeplitz form.* Let Φ be the usual trigonometric system; that is, $\Phi_{2j}(x) := \cos(2\pi jx)$, $\Phi_{2j-1}(x) := \sin(2\pi jx)$, $j = 1, \dots, p$. Define the complex Fourier coefficients of the distribution by setting $d_j := c^{2j} + ic^{2j-1}$, $d_{-j} := c^{2j} - ic^{2j-1}$, $j = 1, \dots, p$ and $d_0 := 1$. We then define the Toeplitz matrix $T(c) := (d_{i-j})_{1 \le i, j \le p}$. Let $h(c) := \det T(c)$. Using the Fejer factorization theorem it is well known [see Krein and Nudel'man (1977), Theorem 2.6, page 65] that the function h satisfies for $c \in \mathscr{K}$ the following properties:

$$\begin{aligned} h(c) = 0 \quad &\Leftrightarrow \quad c \in bd(\mathscr{K}), \\ h(c) \ge 0 \quad &\Leftrightarrow \quad c \in \mathscr{K}. \end{aligned}$$

2.3.2. *Maximum entropy on the mean.* In Gamboa and Gassiat (1991, 1994), we gave a wide family of discriminating functions h. This was the result of constructive methods and stochastic investigation of particular Φ-moment problems. To set them, let us introduce some notation.

1. F is a distribution with support $[0, +\infty]$ such that $F(\{0\}) > 0$.
2. $\psi(t) = \log \int \exp(tx) \cdot dF(x)$ is the log-Laplace transform of F. We assume that ψ has domain $(-\infty, \alpha)$, where $\alpha < +\infty$.
3. For any c in $\mathbb{R}^{2p}$, define

$$h(c) := \alpha - \log F(\{0\}) - \sup_{v \in \mathbb{R}^{2p+1}} \left(\langle v, \tilde{c} \rangle - \int_0^1 \psi(\langle v, \Phi(x) \rangle)\, dx \right),$$

where the angle brackets $(\langle\, , \rangle)$ denote the usual scalar product in $\mathbb{R}^{2p+1}$ and $\tilde{c} := (1, c)$

Interpretation of $h(c)$ in terms of large deviations involving F may be found in Gamboa and Gassiat (1991) and will not be recalled here.

This function h is convex and has the powerful discriminating property that follows [and which is proved in Gamboa and Gassiat (1991)].

PROPOSITION 2.6.

$$\begin{aligned} h(c) = -\infty \quad &\Leftrightarrow \quad c \notin \mathscr{K}, \\ h(c) = 0 \quad &\Leftrightarrow \quad c \in bd(\mathscr{K}), \\ h(c) \ge 0 \quad &\Leftrightarrow \quad c \in \mathscr{K}. \end{aligned}$$

2.3.3. *Hankel forms.* We propose here a similar contrast function using Hankel forms based on the algebraic moments. Let Φ be the sequence of moment functions $\Phi_j(x) = x^j$, $j = 1, \dots, 2p$. Let $M(s)$ be the $(p+1) \times (p+1)$ Hankel matrix given by $M_{i,j} = c_{i+j-2}(s)$, $i, j = 1, \dots, p$, where here $c(s) = E(\Phi(Z_1(s)))$. (Notice that here we do not need to map $\mathbb{R}$ on the interval $(0,1)$.) This matrix is nonnegative as soon as c is the beginning of the moment sequence of a random variable, and degenerates if and only if this random variable is discrete with at most p points of support. If we set $h(c) = \det[M]$ and $H(s) = h(c(s))$, H is a contrast function.

2.4. *Definition and convergence of the estimator.* The sequence H_n is defined as an empirical contrast function in the following way. To use only the observations $Y_1, \dots, Y_n$, we need to truncate the filter s. Let $m(n) < n/2$ be an increasing sequence of integers. Define

$$\hat{Z}(s)_t = \sum_{k=-m(n)}^{+m(n)} s_k Y_{t-k}$$

for $t = 1 + m(n), \dots, n - m(n)$ and

$$c_n(s) := \frac{1}{n - 2m(n)} \sum_{t=1+m(n)}^{n-m(n)} \Phi\big(\varphi\big(\hat{Z}(s)_t\big)\big).$$

We may now define

$$H_n(s) := h\big(c_n(s)\big).$$

We are now able to define our estimator:

DEFINITION 2.7. $\hat{\theta}$ is any minimizer of H_n over Θ_n:

$$\Theta_n = \Theta \cap \{s\colon s_k = 0 \text{ for } |k| > m(n)\}.$$

We assume throughout the sequel that

$$\lim_{n\to\infty} m(n) = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{m(n)}{n} = 0.$$

Our main result is then the following theorem.

THEOREM 2.8. *Assume that* (M1), (M2), (M3) *and* (M4) *hold. If* Θ *is compact, then* $\hat{\theta}$ *converges almost surely, in* $l^1(\mathbb{Z})$, *to* θ *as* n *tends to infinity.*

REMARK 2.1. In practice, we can choose the compact space Θ in the following way:

$$\Theta \subset \left\{ s \in l^1(\mathbb{Z})\colon \sum_k d(k)|s_k| \le M \right\}$$

with $\lim_{k\to\infty} d(k) = +\infty$ and M is a given constant.

REMARK 2.2. Of course, the asymptotic behavior of the estimator depends on the contrast function: The rate depends on the smoothness, and with the same smoothness, the asymptotics differ by the constants involved asymptotically. Further investigation on the asymptotics will give guidelines for the choice of the contrast function.

REMARK 2.3. The use of $(m(n))$ has two simultaneous effects: a truncation of the filtered series and a resulting number of observations involved in the estimation process. To reduce the truncation effect, $m(n)$ should be large, but to increase the number of observations (and consequently the accuracy of the moments estimators) $m(n)$ should be small. A first attempt in the choice of the sequence $[m(n)]$ should be the study of the asymptotic consequences of the two effects, so that a good choice would be a compromise between both. This is similar to the choice of a regularization parameter where bias and variance have to be balanced.

REMARK 2.4. The behavior of the numerical algorithm depends on the choice of the contrast function. We give at the end easy computations for the Toeplitz contrast function. In Gassiat and Gautherat (1994), we give extensive numerical examples using the Hankel contrast function. Other numerical examples using the Hankel contrast function in different models may be found in Dacunha-Castelle and Gassiat (1994) and Gamboa and Gassiat (1995).

3. Some special cases.

In this section, we study the situation where the parameter is finitely parameterized, either with unknown dimension in the AR situation or with known parameterization in the general situation.

3.1. *Autoregressive process.* In this section we will assume that Y is an autoregressive process of unknown order. That is,

$$\theta_k = 0 \quad \text{for } k < -q_1 \text{ and } k > q_2 \text{ for some } q_1, q_2 \in \mathbb{N}.$$

The parameter space is now

$$\Theta := \{s\colon s_0 = 1, \exists q_1, q_2 \in \mathbb{N}, s_k = 0, k < -q_1 \text{ and } k > q_2\}.$$

Notice that here the problem may be seen as purely algebraic: Find a finite filter s that leads to no more than p values for the series $Z_t(s)$. Indeed, as soon as $m(n) > \max(q_1, q_2)$, there are no truncation effects. However, numerical algorithms for such algebraic guesses are not evident, so that our optimization method stays interesting. Moreover, its speed of convergence is exactly the same as the best algebraic one. We shall use the following assumption:

(M5) For all $a \in l^1(\mathbb{Z})$ such that $C(a) := \{j\colon a_j \neq 0\}$ is infinite, the distribution of $W(a)$ is continuous.

THEOREM 3.1. *Under assumptions* (M1), (M2) *and* (M5), *if* $\hat{\theta}$ *is chosen as the minimum length minimizer of* H_n *in* Θ_n, *then*

$$\text{as soon as } n - 2m(n) > p + 1 \text{ and } m(n) > \max(q_1, q_2), \qquad \hat{\theta} = \theta.$$

Notice that we do not assume here that the input process is stationary. The only assumption made is that any infinite linear combination built with X has continuous distribution. This is true, for example, if X is an independent sequence and

$$\exists \varepsilon > 0, \qquad P(X_i = x_j) \geq \varepsilon, \qquad i \in \mathbb{N}, j = 1, \dots, p.$$

Indeed, in this case, applying theorems of Jessen Wintner [Theorem VI.23.4 in Hennequin and Tortrat (1965)] and of Lévy [Theorem VI.23.5 in Hennequin and Tortrat (1965)], the distribution of $W(a)$ is pure and continuous as soon as $C(a)$ is infinite. The proof of Theorem 3.1 relies on the following finite sample property of $\hat{Z}(s)$:

PROPOSITION 3.2. *Assume* (M1), (M2) *and* (M5). *Then*, *for* $n - 2m(n) > p$,

$$P\Big(\#\big\{\hat{Z}(s)_t, t = 1 + m, \dots, n - m\big\} \leq p\Big) > 0$$

implies that s *is a finite linear combination of shifted* θ.

We conjecture that Theorem 3.1 stays true with much weaker conditions than (M5). Using similar arguments as in Proposition 2.1, we believe that it can be shown with extra work that $\hat{\theta}$ converges almost surely to θ in a finite number of steps. Indeed, looking at the proof of Proposition 3.2, we see that $\hat{\theta} = \theta$ a.s. as soon as

$$\#\big\{\hat{Z}(s)_t, t = 1 + m, \dots, n - m\big\} > p \quad \text{if } s \neq \theta.$$

EXAMPLE. AR1 Process. As an example let us consider the simplest model:

$$Y_t - \theta Y_{t-1} = X_t, \qquad |\theta| \neq 1, t \in \mathbb{Z}.$$

Recall that the estimator is perfect when the order of the AR is unknown. However, if we know that the model is AR1, the process $Z(s)$ has the simple form

$$Z_t(s) = \hat{Z}(s)_t = Y_t - sY_{t-1}, \qquad t = 2, \dots, n,$$

and $c_n(s) = (n-1)^{-1}\sum_{t=2}^{n} \Phi(\varphi(\hat{Z}(s)_t))$. Theorem 3.1 says that for $n > p + 2$, θ can be found exactly. With binary inputs the contrast function given in Section 2.3 (Toeplitz) is

$$\begin{aligned} h(c) = 1 &- 2\Big((c^1)^2 + (c^2)^2\Big) - \Big((c^3)^2 + (c^4)^2\Big) \\ &+ 2\Big((c^1)^2 c^3 + 2c^1c^2c^4 - (c^2)^2 c^3\Big). \end{aligned}$$

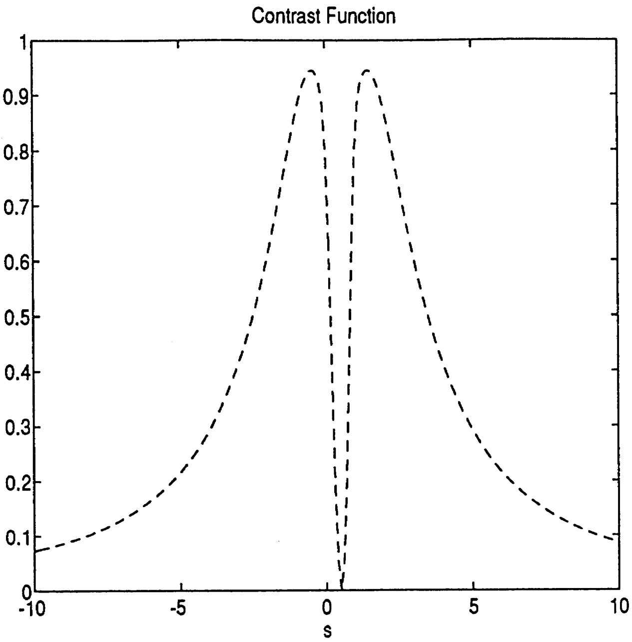

FIG. 1. *Contrast function using Toeplitz determinant.*

The function $H(s) = h(c(s))$ is not easily tractable. Its graph is drawn in Figure 1 for $\theta = \frac{1}{2}$. A realization of the graph of H_n is drawn in Figure 2 for various values of $n(\varphi(x) := (1/\pi)\arctan x + \frac{1}{2})$.

3.2. *The parametric case.* Suppose that the set Θ can be represented as a parametric model with real-valued parameter vector ζ in a set $\mathscr{S}$ of dimension l: $\zeta = (\zeta_j)_{j=1,\dots,l}$:

$$\Theta := \{s(\zeta), \zeta \in \mathscr{S}\}.$$

Let ζ^* be the true parameter value. To estimate ζ^*, we minimize $L_n(\zeta) := H_n(s(\zeta))$. Let $\hat{\zeta}$ be any minimizer of L_n over a given compact set K containing ζ^*.

COROLLARY 3.3. *Assume that the application $\zeta \to s(\zeta)$ from $\mathbb{R}^l$ to $l^1(\mathbb{Z})$ is continuous and that assumptions* (M1)–(M4) *hold. Assume the identifiability assumption*:

$$s_k(\zeta) = rs_{k-K}(\zeta') \qquad \forall k \in \mathbb{Z} \Leftrightarrow r = 1, K = 0 \text{ and } \zeta = \zeta'.$$

Then $\hat{\zeta}$ converges, almost surely, as n approaches infinity, to ζ^*.

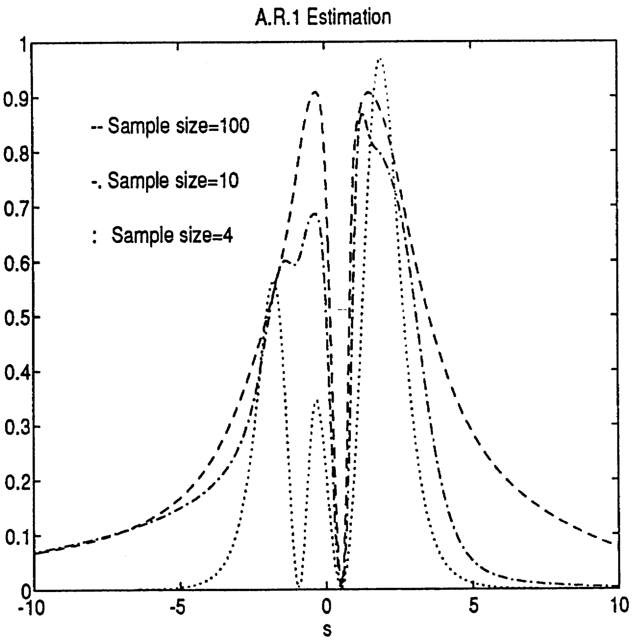

FIG. 2. *Estimation for the model* $Y_t - \theta Y_{t-1} = X_t$.

REMARK 3.1. In the case of Hankel forms and under some smoothness assumptions on the mapping $\zeta \to s(\zeta)$, the following inequality is proved in Gassiat and Gautherat (1995): There exists a constant $K > 0$ such that for n large enough,

$$\|\hat{\zeta} - \zeta^*\| \leq K \sum_{|j| > m(n)} |s(\zeta^*)| \quad \text{(a.s.)} \tag{3.1}$$

(here, $\|\cdot\|$ denotes a norm on $\mathbb{R}^l$ and the constant K depends on ζ^*).

REMARK 3.2. Simulations in Gassiat and Gautherat (1994) lead to apparently perfect estimation in MA models. This experimental result may be explained by (3.1). Indeed, in this case the right side of (3.1) goes to zero exponentially fast.

4. Simultaneous estimation of the cardinality p. In this section, we assume that the cardinality p of the support A is unknown. No upper bound on p has to be known. For any possible cardinality $q > 0$ we consider the contrast function built *as if q were the true cardinality of the support*. Let $H(s, q) := h_q(c^q(s))$, $s \in \Theta$, where

$$c^q(s) := E\left[\Phi^{(q)}(\varphi(Z_1(s)))\right],$$
$$\Phi^{(q)} := (\Phi_1, \dots, \Phi_{2q});$$

$(\Phi^{(q)})_{q \geq 1}$ is such that for all q, $(1, \Phi_1, \dots, \Phi_{2q})$ is a T-system. Such systems of functions are called M-systems; see Krein and Nudel'man (1977). For instance, trigonometric or algebraic functions as described in Sections 2.3.1 and 2.3.3 are M-systems. h_q is a given nonnegative function defined on $\mathscr{K}_q$, the set of $\Phi^{(q)}$ moment sequences of probabilities on $(0, 1)$ that only vanishes on the boundary of $\mathscr{K}_q$. We have the following characterization of the parameters, which is an obvious consequence of Theorem 2.3:

THEOREM 4.1. (i) *If* $q < p$, *then* $\forall s \in \Theta$, $H(s, q) > 0$.
(ii) $H(\theta, p) = 0$ *and* $\forall s \in \Theta$, $s \neq \theta$, $H(s, p) > 0$.
In other words, p is the smallest integer q such that the equation $H(s, q) = 0$ has a solution s, which is then θ.

To estimate (θ, p), we consider the empirical contrast functions $H_n(s, q)$. Minimizing simultaneously in both variables s and q would not lead to good estimators; it would lead to systematic overestimation of p. To rank small values of the integer q, we shall use a compensation technique, as usual for the estimation of the order of a model; see, for instance, Azencott and Dacunha-Castelle (1984) or Dacunha-Castelle and Gassiat (1994). Let $\delta(n)$ be a sequence of positive real numbers with limit 0 as n tends to infinity. Define

$$J_n(s, q) = H_n(s, q) + \delta(n) \cdot q.$$

Let now the estimator $(\hat{\theta}, \hat{p})$ be defined as a minimizer of J_n over $\Theta_n \times \mathbb{N}^*$. To have good asymptotic properties of the estimator, roughly speaking, the compensation sequence $\delta(n)$ has to be related to the stochastic variations of H_n, so that we introduce the following assumption:

(M6) Let $\check{\theta}$ be the truncation of the filter θ [$\check{\theta}_k = \theta_k$ if $|k| \leq m(n)$ and $\check{\theta}_k = 0$ if $|k| > m(n)$]:

$$\lim_{n \to \infty} \frac{H_n(\check{\theta}, p)}{\delta(n)} = 0 \quad \text{a.s.}$$

We then have the following theorem:

THEOREM 4.2. *Assume that Θ is a compact subset of $l^1(\mathbb{Z})$. Under assumptions* (M1), (M2), (M3), (M4) *and* (M6), *as n tends to infinity, $\hat{\theta}$ converges a.s. in l_1 to θ and $\hat{p}$ converges a.s. to p.*

REMARK ON ASSUMPTION (M6). As soon as the discriminating function has continuous derivatives (which is the case for Toeplitz and Hankel forms), assumption (M6) relates the speed of convergence of the Φ-moments to the compensator sequence $\delta(n)$. For instance, (M6) holds with $\delta(n)^{-1} = o(\sqrt{n/(\log\log n)})$ as soon as the input sequence obeys the iterated logarithm law.

5. Extension to other models.

5.1. *Nonstationary inputs.* Roughly speaking, the method proposed in this paper works because of the properties of the empirical measure $\nu_n(s)$, built on the observations

$$\nu_n(s) := \frac{1}{n-2m(n)} \sum_{t=1+m(n)}^{n-m(n)} \delta_{\varphi(\hat{Z}(s)_t)}.$$

By ergodicity, this sequence of random measures converges weakly to a distribution which has support of cardinal p if and only if $s = \theta$. Now, define the random measure without truncation $\tilde{\nu}_n(s)$:

$$\tilde{\nu}_n(s) := \frac{1}{n-2m(n)} \sum_{t=1+m(n)}^{n-m(n)} \delta_{\varphi(Z_t(s))}.$$

Without any assumption on the probabilistic nature of the input signal, it is easy to see that $\tilde{\nu}_n(s)$ has, for any n and for $s = \theta$, at most p points of support. This applies with increasing probability only for $s = \theta$ under very wide assumptions, for instance, under (M4) for $n > p$. Now, in the nonstationary case, if this remains true for all the accumulation points of the sequence $(\nu_n(s))$, all the previous results stay valid.

5.2. *Multidimensional systems.* Similar ideas may be developed in the case where the variables X_t are random vectors (i.e., are multidimensional) by an appropriate choice of the functions Φ_j. In this case, difficulties arise from the fact that T-systems of functions do not exist [see Krein and Nudel'man (1977), page 32] so that an analogue of Theorem 2.3 is not obvious. However, Proposition 2.1 and Theorem 2.2 may be easily extended (just using some partial order in multidimensional space), so that the structure of the model and the characterization of the inverse filter is essentially the same as in the one-dimensional situation.

6. Proofs.

PROOF OF PROPOSITION 2.1. Without loss of generality, we may assume that the integers 1 and 2 are in $C(a)$. Using assumption (M4), the support of the variable $a_1X_1 + a_2X_2$ is exactly $\{a_1x_i + a_2x_j,\ i, j = 1,\dots,p\}$, which contains at least $p+1$ distinct points $\xi_1 < \cdots < \xi_{p+1}$. Now:

(i) If $C(a)$ is finite [using again assumption (M4)], we may deduce that the support of $W(a)$ contains the following distinct $p+1$ points:

$$\xi_j + x_1 \textstyle\sum_{l \in C(a),\, l \neq 1,2} a_l,\ j = 1,\dots,p+1.$$

(ii) If $C(a)$ is infinite, let

$$W_N(a) = \sum_{|j|>N} a_j X_j.$$

The following inequality holds almost surely:

$$|W_N(a)| \le \sup\{|x_1|,\dots,|x_p|\} \sum_{|j|>N} |a_j|. \tag{6.1}$$

Let $\varepsilon < \inf\{\xi_{j+1} - \zeta_j\}$, $j = 1,\dots,p$, and N be such that

$$\sup\{|x_1|,\dots,|x_p|\} \sum_{|j|>N} |a_j| < \varepsilon.$$

Define now

$$I_j = \left(\xi_j + x_1 \sum_{|l|\le N,\, l\ne 1,2} a_l - \varepsilon,\ \xi_j + x_1 \sum_{|l|\le N,\, l\ne 1,2} a_l + \varepsilon \right).$$

The intervals I_j are disconnected thanks to the choice of ε, and $P(W(a) \in I_j) > 0$ thanks to assumption (M4), the bound (6.1) and the choice of ε. □

Proof of Theorem 2.5. By definition, $c(s) = E[\phi(\varphi(Z_1(s)))]$. Using Theorem 2.3, $c(s)$ lies on the boundary of $\mathscr{K}$ if and only if the distribution of $\varphi(Z_1(s))$ is purely atomic with support of cardinality less than or equal to p. However, $Z_1(s) = \sum_{k\in\mathbb{Z}} (s*u)_k \cdot X_{1-k}$ so that, applying Proposition 2.1, only one $(s*u)_k$ can be nonzero. That is, $\forall t \in \mathbb{Z}$, $s_t = \theta_t$. □

Proof of Theorem 2.8. As the sequence $\hat\theta$ belongs to a compact set, it possesses at least one accumulation point $\bar\theta$. Now

$$\forall s \in \Theta, \qquad H_n(\hat\theta) \le H_n(s),$$

so that

$$\limsup_{n\to\infty} H_n(\hat\theta) \le H(s) \quad \text{a.s.} \tag{6.2}$$

Indeed, h is a continuous function on $\mathscr{K}$, and by Lemma 6.2,

$$\forall s \in \Theta\colon \lim_{n\to\infty} c_n(s) = c(s).$$

In this proof, all the involved limits are almost sure so that the convergence of $\hat\theta$ will be almost sure.

We now prove that

$$\lim_{n\to\infty} H_n(\hat\theta) = H(\bar\theta)$$

so that (6.2) together with Theorem 2.5 implies $\hat\theta = \theta$ and, consequently, the almost sure convergence of θ to θ:

$$H_n(\hat\theta) - H(\bar\theta) = H_n(\hat\theta) - H_n(\bar\theta) + H_n(\bar\theta) - H(\bar\theta).$$

Now

$$\lim_{n\to\infty} H_n(\bar\theta) - H(\bar\theta) = 0,$$

$$H_n(\hat\theta) - H_n(\bar\theta) = h\big(c_n(\hat\theta)\big) - h\big(c_n(\bar\theta)\big).$$

For any filters s_1 and s_2,

$$c_n(s_1) - c_n(s_2) \tag{6.3}$$
$$= \frac{1}{n - 2m(n)} \sum_{t=1+m(n)}^{n-m(n)} \left[\Phi\big(\varphi(Z_t(s_1))\big) - \Phi\big(\varphi(Z_t(s_2))\big)\right].$$

$\Phi \circ \varphi$ is uniformly continuous on $\mathbb{R}$, so that using (6.3) and Lemma 6.1, $\|c_n(s_1) - c_n(s_2)\|$ becomes arbitrarily small with $\|s_1 - s_2\|_1$.

Now $\mathscr{K}$ is a compact set, so that the function h is uniformly continuous on $\mathscr{K}$. So $|h(c_n(s_1)) - h(c_n(s_2))|$ becomes arbitrarily small with $\|s_1 - s_2\|_1$. Therefore, since $\bar{\theta}$ is an accumulation point of $(\hat{\theta})$, we find

$$\lim_{n\to\infty} H_n(\hat{\theta}) = H(\bar{\theta})$$

and the proof is complete. □

LEMMA 6.1.

$$\forall s_1, s_2 \in l^1(\mathbb{Z}), \qquad |Z_t(s_1) - Z_t(s_2)| \le \sup\big\{|x_1|, \dots, |x_p|\big\}\|u\|_1\|s_1 - s_2\|_1.$$

PROOF.

$$Z_t(s_1) - Z_t(s_2) = \sum_k \left[(u * s_1)_k - (u * s_2)_k\right] X_{t-k},$$

so that, $\forall t$,

$$\begin{aligned} |Z_t(s_1) - Z_t(s_2)| &\le \|u * (s_1 - s_2)\|_1 \|X_1\|_\infty \\ &\le \sup\big\{|x_1|, \dots, |x_p|\big\}\|u\|_1\|s_1 - s_2\|_1. \end{aligned}$$ □

LEMMA 6.2.

$$\forall s \in \Theta\colon \lim_{n\to\infty} c_n(s) = c(s) \quad a.s.$$

PROOF. Let, for $s \in \Theta$,

$$\tilde{c}_n(s) := \frac{1}{n} \sum_{t=1}^{n} \Phi\big(\varphi\big(\hat{Z}(s)_t\big)\big),$$
$$\hat{c}_n(s) := \frac{1}{n} \sum_{t=1}^{n} \Phi\big(\varphi(Z_t(s))\big).$$

Thanks to the ergodic theorem, $(\hat{c}_n(s))$ converges almost surely to $c(s)$. From Lemma 6.1 this is the same for $(\tilde{c}_n(s))$. Now

$$c_n(s) = \frac{n}{n - 2m(n)}\tilde{c}_n(s) - \frac{1}{n - 2m(n)} \sum_{t=1}^{m(n)} \Phi\big(\varphi\big(\hat{Z}(s)_t\big)\big).$$

Therefore, as $m(n)$ has been chosen such that $\lim_{n\to\infty}(m(n))/n = 0$, $(c_n(s))$ converges almost surely to $c(s)$. □

PROOF OF PROPOSITION 3.2.

LEMMA 6.3. *The set* $\{k \in \mathbb{Z}, (u * s)_k \neq 0\}$ *has finite cardinality if and only if*

$$\exists h \in \mathbb{N}, \exists(a_1, \dots, a_h) \in \mathbb{R}^h, \exists(l_1, \dots, l_h) \in \mathbb{Z}^h: \forall k, \qquad s_k = \sum_{i=1}^{h} a_i \theta_{k-l_i}.$$

PROOF. Obvious, with $\{k \in \mathbb{Z}, (u * s)_k \neq 0\} = \{l_1, \dots, l_h\}$ and $(u * s)_{l_i} = a_i$. □

We have, as soon as $n - 2m(n) > p$,

$$\begin{aligned} P\Big(\#\big\{\hat{Z}(s)_t, t = 1 + m, \dots, n - m\big\} \leq p\Big) &\leq \sum_{1+m \leq t_i \neq t_j \leq n-m} P\Big(\hat{Z}_{t_i}(s) = \hat{Z}_{t_j}(s)\Big) \\ &= \sum_{1+m \leq t_i \leq t_j \leq n-m} P\Big(\hat{Z}_{t_i}(s) - \hat{Z}_{t_j}(s) = 0\Big). \end{aligned}$$

For $j \in \mathbb{Z}$, let $\tilde{s}_j := s_j 1_{|j| \leq m(n)}$ (the truncated filter). Now,

$$\hat{Z}_{t_i}(s) - \hat{Z}_{t_j}(s) = \sum_k \big[(u * \tilde{s})_{t_i - k} - (u * \tilde{s})_{t_j - k}\big] X_k$$

and the set

$$\big\{k \in \mathbb{Z}: (u * \tilde{s})_{t_i-k} - (u * \tilde{s})_{t_j-k} \neq 0\big\} = \big\{k \in \mathbb{Z}: (u * \tilde{s})_{t_i-k} \neq (u * \tilde{s})_{t_j-k}\big\}$$

has infinite cardinality as soon as the set

$$\{k \in \mathbb{Z}: (u * \tilde{s})_k \neq 0\}$$

has infinite cardinality too. Indeed, if it was not, there would exist an integer k_0 such that

$$\forall k \geq k_0, \qquad (u * \tilde{s})_k = (u * \tilde{s})_{k+t_j-t_i}$$

and the sequence $((u * \tilde{s})_k)_{k \geq k_0}$ would be periodic with period $|t_j - t_i|$, which is impossible since it is summable, except if it is identically 0.

Then, using assumption (M5), we may conclude that if the set $\{k \in \mathbb{Z}: (u * \tilde{s})_k \neq 0\}$ is infinite, then the distribution of $\hat{Z}_{t_i}(s) - \hat{Z}_{t_j}(s)$ is continuous for all distinct t_i and t_j. Consequently, $P(\hat{Z}_{t_i}(s) = \hat{Z}_{t_j}(s)) = 0$ and $P(\#\{\hat{Z}(s)_t, t = 1 + m, \dots, n - m\} \leq p) = 0$. By Lemma 6.3 the set $\{k \in \mathbb{Z}: (u * \tilde{s})_k \neq 0\}$ is finite if and only if

$$\begin{aligned} &\exists h \in \mathbb{N}, \exists a_1, \dots, a_h \in \mathbb{R}, \exists l_1, \dots, l_h \in \mathbb{Z}, \forall k \in \mathbb{Z}: \\ &\tilde{s}_k = \sum_{i=1}^{h} a_i \theta_{k-l_i}. \end{aligned} \tag{6.4}$$

□

PROOF OF THEOREM 3.1. If $s_k = \sum_{i=1}^{h} a_i \theta_{k-l_i}$ with at least two nonzero coefficients a_i, then the length of s is greater than the length of θ. □

PROOF OF THEOREM 4.2. Since Θ is compact in $l_1(\mathbb{Z})$, the sequence $(\hat{\theta})$ possesses at least one a.s. accumulation point θ^*. If this accumulation point is proved to be unique and equal to θ and if $\hat{p}$ is proved to converge to p, then the theorem is proved. We now work on the subsequence converging to θ^*. Using the definition of the estimator, we have

$$J_n(\hat{\theta}, \hat{p}) \le J_n(\bar{\theta}, p).$$

We then have

$$\frac{H_n(\hat{\theta}, \hat{p})}{\delta(n)} + \hat{p} \le \frac{H_n(\bar{\theta}, p)}{\delta(n)} + p.$$

Since $H_n(\hat{\theta}, \hat{p})$ and $\delta(n)$ are nonnegative, this implies

$$\hat{p} \le \frac{H_n(\bar{\theta}, p)}{\delta(n)} + p,$$

so $\limsup_{n\to\infty} \hat{p} \le p$ and the sequence $\hat{p}$ is bounded. Let p^* be an accumulation point of the sequence $\hat{p}$. We thus have

$$p^* \le p \quad \text{a.s.} \tag{6.5}$$

using (M6). Now, using the same tricks as for the proof of Theorem 2.8, it can be proved that

$$\lim_{n\to\infty} H_n(\hat{\theta}, \hat{p}) = H(\theta^*, p^*) \quad \text{a.s.} \tag{6.6}$$

We then obtain

$$H(\theta^*, p^*) = 0. \tag{6.7}$$

Now, using Theorem 4.1, (6.5) and (6.7) imply

$$p^* = p, \qquad \theta^* = \theta$$

and the theorem is proved. □

REFERENCES

AZENCOTT, R. and DACUNHA-CASTELLE, D. (1984). *Séries d'Observations Irrégulières, Modélisation et Prévision. Techniques Stochastiques*. Masson, Paris.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1994). Estimation of the number of components in a mixture. *Bernoulli*. To appear.

FEHER, K., ed. (1987). *Advanced Digital Communications—Systems and Signal Processing Techniques*. Prentice-Hall, Englewood Cliffs, NJ.

GAMBOA, F. and GASSIAT, E. (1991). Bayesian methods and maximum entropy for ill posed inverse problems. Unpublished manuscript.

GAMBOA, F. and GASSIAT, E. (1994). The maximum entropy method on the mean: applications to linear programming and superresolution. *Math. Programming* **66** 103–122.

GAMBOA, F. and GASSIAT, E. (1995). Source separation when the input sources are discrete or have constant modulus. Unpublished manuscript.

GASSIAT, E. (1990). Estimation semi-paramétrique d'un modèle autorégressif stationnaire multi-indice non nécessairement causal. *Ann. Inst. H. Poincaré Probab. Statist.* **26** 181–205.

GASSIAT, E. (1993). Adaptive estimation in noncausal AR processes. *Ann. Statist.* **21** 2022–2042.

GASSIAT, E. and GAUTHERAT, E. (1994). Identification of noisy linear systems with discrete random input. *IEEE Trans. Inform. Theory*. To appear.

GASSIAT, E. and GAUTHERAT, E. (1995). Some rates in blind discrete deconvolution. Unpublished manuscript.

HENNEQUIN, P. and TORTRAT, A. (1965). *Théorie des Probabilités et Quelques Applications*. Masson, Paris.

KREIN, M. G. and NUDEL'MAN, A. A. (1977). The Markov moment problem and extremal problems. *Bull. Amer. Math. Soc.* **50**.

KREISS, J. P. (1987). On adaptive estimation in stationary A.R.M.A. processes. *Ann. Statist.* **15** 112–133.

LI, T. H. (1992). Blind identification and deconvolution of linear systems driven by binary random sequences. *IEEE Trans. Inform. Theory* **38** 26–38.

LI, T. H. (1995). Blind deconvolution of linear systems with multilevel nonstationary inputs. *Ann. Statist.* **23** 690–704.

LII, K. S. and ROSENBLATT, M. (1982). Deconvolution and estimation of transfer function phase and coefficients for non-Gaussian linear processes. *Ann. Statist.* **10** 1195–1208.

INSTITUT GALILÉE
UNIVERSITÉ PARIS-NORD
F-93430 VILLETANEUSE
FRANCE
E-MAIL: gamboa@stats.matups.fr

LABORATOIRE ANALYSE ET PROBABILITÉ
UNIVERSITÉ D'EVRY VAL D'ESSONNE
BOULEVARD DES COQUIBUS
F-91025 EVRY CEDEX
FRANCE
E-MAIL: gassiat@stats.matups.fr