

SPARSE REPRESENTATIONS AND REALIZATION THEORY

Jean-Jacques FUCHS

IRISA/Université de Rennes I
Campus de Beaulieu - 35042 Rennes Cedex - France
fuchs@irisa.fr

ABSTRACT

We present some results obtained recently in signal processing in the so-called “sparse representations” domain and indicate how they can be applied to a very specific and limited problem in realization theory. This is mainly to bring these type of results to the knowledge of this community. Other applications in order estimation for instance are potentially feasible. The basic problem is the following: given a (n, m) -matrix A with $m > n$ and a vector $b = Ax_o$ with x_o having p non-zero components, find sufficient conditions for x_o to be the unique sparsest solution of $Ax = b$, the answer is an upper-bound on p depending upon A . We present as application the realization of a partial covariance sequences.

1. INTRODUCTION

Sparse approximation is the problem of finding a representation of a signal or a function as a linear combination of a small number of elements from an over-complete set of vectors, signals or functions often called a dictionary or a redundant basis. Indeed several problems are of interest depending on the context. One may seek the sparsest exact representation of a signal in terms of the elements or the representation of a given complexity that minimizes a certain approximation error or the sparsest representation that yields an approximation error smaller than a specified threshold.

Some results concerning the first of these questions have been obtained recently. Given an (n, m) -matrix A with $m > n$ and a vector b that indeed admits an exact sparse representation, say $b = Ax_o$, it has been shown that if the number of non-zero entries in x_o is smaller than a given bound, then x_o is the unique sparsest representation.

Since searching for the sparsest representation is a non-polynomial (NP) hard problem [1] that can only be solved by exhaustive search, one is tempted to replace the true search for the sparsest solution :

$$\min_x \|x\|_0 \quad \text{subject to} \quad Ax = b \quad (P_0)$$

with $\|x\|_0$ the number of non-zero entries in x , by the following easy-to-solve optimization problem :

$$\min_x \|x\|_1 \quad \text{subject to} \quad Ax = b \quad (LP)$$

i.e., to minimize the ℓ_1 norm of x instead of the sparsity itself. Here and below $\|x\|_k$ denotes the ℓ_k norm of a vector x , defined as $\|x\|_k = \{\sum_1^m |x_j|^k\}^{1/k}$ for $k \geq 1$. The problem is then to determine sufficient conditions for the two criteria to have the same unique solution. This problem has been initiated in [2] and developed since then by several other authors, e.g., [3, 4, 5].

Notice that we denote this optimization problem by (LP) though it is not a linear program in standard form. It is of course easy to put it in the standard form and since its optimum is then always (also) attained at a basic feasible solution to deduce that there is an optimum that has at most n non-zero components. Here assuming that there exists a far sparser admissible point x_o , one seeks conditions under which x_o is the (degenerate) optimum of (LP) [6].

Sufficient conditions under which this is the case admit different expressions according to the assumptions one is willing to make. If the emphasis is on sparsity one generally normalizes the columns a_j of A to one in Euclidean norm and introduces the mutual coherence [2]

$$M = \max_{1 \leq i \neq j \leq m} |a_i^T a_j|, \quad (1)$$

of the dictionary whose components are the columns of A . The smaller M , the less coherent are the components of the dictionary and $M = 0$ if and only if the columns are orthogonal. It is worthwhile to know that there are indeed redundant dictionaries with $m \simeq n^2$ components and mutual coherence $M \simeq \frac{1}{\sqrt{n}}$ [8, 9].

It is shown in [3, 4, 5] that if

$$\|x_o\|_0 < \frac{1}{2} \left(1 + \frac{1}{M}\right) \quad (2)$$

then x_o is the unique sparsest representation of b that can furthermore be recovered by solving (LP) with $b = Ax_o$.

It is worth noting that (2) is independent of the magnitudes of the nonzero entries of x_o . Being able to recover x_o

appears to be only a matter of structure, of angles between vectors. This is similar to what happens with identifiability conditions in estimation theory or observability conditions in systems theory. It is only in the presence of noise, when $b = Ax_o + e$ with e a vector of white Gaussian noise for instance that the magnitudes of the non-zero components in x_o come into play [7].

2. AN APPLICATION IN REALIZATION THEORY

Here we will consider a very specific A matrix and restrict our attention to the case where the weights in x_o used to build the observed vector $b = Ax_o$ are known to be greater than or equal to zero. The sign restriction on the weights changes quite drastically the nature of the problem and the condition under which recovery is possible we will obtain will be quite different from (2).

We will consider the case where the A matrix is a Vandermonde matrix whose columns, we denote $v(\alpha)$, are of the form

$$v(\alpha)^T = [1 \ \alpha \ \alpha^2 \ \dots \ \alpha^k \ \dots \ \alpha^{n-1}], \quad |\alpha| < 1 \quad (3)$$

The reason for considering solving (LP) with such an A matrix is that it is a preliminary step towards the realization of a partial covariance sequence of a stationary time series of finite order.

Indeed if b contains the exact partial covariance sequence of a stationary stochastic process that is an order- p autoregressive (AR(p)) process which is further known to have only distinct real poles belonging to a known grid in the interval $] -1, 1[$, then

$$b = \sum_{j=1}^p x_{i_j} v(\alpha_{i_j})$$

and we will find the conditions that guarantee that the optimum of the sign-restricted (LP) yields precisely the minimal realization of this partial covariance sequence.

In section 3, we collect all the results from linear programming theory we need in the sequel, in section 4 we get the conditions under which the sought-for the solution is the unique solution to (LP). In section 5 we explain the links between the results obtained in section 4 and realization theory, we present some further remarks in section 6 and conclude in section 7.

3. OPTIMALITY CONDITIONS FOR (LP)

Since we assume the weights in x to be greater than or equal to zero, the linear program (LP) minimizing the ℓ_1 -norm becomes

$$\min_x \mathbf{1}^T x \quad \text{subject to} \quad Ax = b \quad \text{and} \quad x \geq 0 \quad (\text{LP+})$$

where $\mathbf{1}$ denotes a column vector of ones of adequate dimension. To get the optimality conditions in a convenient form we introduce the dual of this linear program [6]:

$$\max_d b^T d \quad \text{subject to} \quad A^T d \leq \mathbf{1} \quad (\text{DLP+})$$

For given b , let $x_o \geq 0$ be a feasible point of (LP+), we denote \bar{x}_o the reduced dimensional vector built with the strictly positive components in x_o and \bar{A}_o the matrix built with the corresponding columns of A , one thus has, e.g., $Ax_o = \bar{A}_o \bar{x}_o = b$. Standard duality theory for convex programs then says that the point x_o is an optimum of (LP+) if and only if there exists, say d_o , a dual feasible point that achieves the same cost, i.e.

$$\exists d_o \ni A^T d_o \leq \mathbf{1} \quad \text{and} \quad \mathbf{1}^T x_o = b^T d_o$$

This point d_o is then an optimum of (DLP+). Furthermore both optima are unique if no constraint of the dual is degenerate, i.e., $a_j^T d_o < 1 \ \forall a_j \notin \bar{A}_o$. To summarize one has the following proposition.

Lemma 1: The point x_o is the unique optimum of (LP+) if

$$\begin{aligned} \exists d_o \ni \quad & \mathbf{1}^T x_o = b^T d_o, \quad \bar{A}_o^T d_o = \mathbf{1} \\ & \text{and} \quad a_j^T d_o < 1, \quad \forall a_j \notin \bar{A}_o \quad \square \end{aligned} \quad (4)$$

This is only a sufficient condition for uniqueness but since the solution of the primal linear program (LP+) is degenerate, i.e. $\|x_o\|_0 < n$, one cannot expect a stronger result. When the optimum of the primal is degenerate the optimum of the dual (DLP+) is undetermined and this also signifies that the primal is *unstable*, i.e., even the slightest perturbation of b will lead to a drastic change in the optimum of the primal. The optimum of the primal will loose its sparsity in a fully unpredictable way.

In the sequel we will apply Lemma 1 to the specific A matrix whose columns are given by (3). In [4], it is shown that when applied to arbitrary matrices with normalized columns, the conditions of the lemma become condition (2). We will not assume the columns in A to be normalized.

4. SUFFICIENT CONDITION FOR RECOVERY

The (n,m) -matrix A has m columns denoted $v(\alpha_i)$ defined in (3) with α_i having m distinct values taken for instance on a regular grid in $] -1, 1[$. The matrix A is known as a Vandermonde matrix. We seek sufficient conditions on $\|x_o\|_0$, under which the solution of (LP+) with $b = Ax_o$ is unique and equal to x_o which is assumed to satisfy $x_o \geq 0$. It happens that the conditions we will get are independent of m and on the way the m distinct values α_i are taken. They only depend upon n the number of observations. We prove

the following results.

Lemma 2: For A an (n,m) -Vandermonde-matrix with distinct columns, the sparsest solution to $\{Ax = b, x \geq 0\}$ and to (LP+) is unique if $b = Ax_o$ with $x_o \geq 0$ and

$$\|x_o\|_0 \leq \lfloor \frac{n-1}{2} \rfloor \quad (5)$$

where $\lfloor z \rfloor$ denotes the nearest integer that is $\leq z$. \square

Proof: We indicate how to construct a vector d_o satisfying (4). With an arbitrary vector $f = [f_0 \ f_1 \ \dots \ f_{n-1}]^T$ we associate the polynomial

$$F(\alpha) = f^T v(\alpha) = f_0 + f_1 \alpha + \dots + f_{n-1} \alpha^{n-1}$$

The idea is to build a polynomial in α that is zero on the values α_k to be preserved and strictly positive elsewhere.

With a given α_k , one associates $P_{\alpha_k}(\alpha) = -\alpha_k + \alpha$ whose square: $P_{\alpha_k}^2(\alpha) = \alpha_k^2 - 2\alpha_k \alpha + \alpha^2$ is then such that $P_{\alpha_k}^2(\alpha) > 0, \forall \alpha \neq \alpha_k$ and $P_{\alpha_k}^2(\alpha_k) = 0$. But $P_{\alpha_k}^2(\alpha) = f_k^T v(\alpha)$ with

$$f_k^T = [\alpha_k^2 \ -2\alpha_k \ 1 \ 0 \ 0 \ \dots \ 0]$$

and the vector $d^T = [1 \ 0 \ 0 \ 0 \ \dots \ 0] - f_k^T$ then satisfies conditions (4) if $\bar{A}_o = [v(\alpha_k)]$.

If there are p columns in \bar{A}_o , one proceeds similarly and one builds the polynomial $P^2(\alpha) = \prod_{k=1}^p P_{\alpha_k}^2(\alpha)$. This polynomial has degree $2p$, the associated vector f has $2p + 1$ non-zero components and so has the sought for vector d . The bound (5) then follows by writing that the number of non-zero components in d has to be smaller than or equal to the dimension n of d . \square

Note that the bound (5) does not depend on m the number of columns in A and holds for both un-normalized or normalized $v(\alpha)$ columns (3) since the proof remains valid if the columns are normalized.

To fix ideas, we assumed that the α_i 's are equispaced in $]-1, 1[$ but this is nowhere used in the proof and can be relaxed. On the opposite, the assumption on the sign of the weights ($x \geq 0$) is essential in our proof and seems necessary if one wants to improve the bound from (2) to (5) in the linear programming context.

5. COMMENTS AND LINKS WITH PARTIAL REALIZATION THEORY

The bound (5) we obtained, is the one that guarantees that solving the linear program (LP) will yield the *good* solution.

The necessary and sufficient bound on $\|x_o\|_0$ for x_o to be the unique sparsest solution to $Ax = Ax_o$ (true sparsity and not just ℓ_1 -sparsity) is easy to obtain in this case. Since for (n, m) -dimensional Vandermonde matrices with distinct columns all sets of n column-vectors are linearly independent, the vectors in the kernel of A have at least $n + 1$ non-zero components and it follows that the expected bound is

$$\|x_o\|_0 \leq \lfloor \frac{n}{2} \rfloor$$

It is identical to (5) for odd n and slightly better for even n . This may be the price to pay (in addition to the sign constraint) to recover x_o without an exhaustive combinatorial search.

The bound (5) is also the one, one obtains from the *realization theory* of a partial covariance sequence[10]. The components in b are assumed to be the first elements of the covariance sequence of a stationary autoregressive process of order say p (AR(p)) having single real poles. The square Hankel matrix built with the n samples of the covariance sequence in b is of maximal order $\lfloor \frac{n+1}{2} \rfloor$ and it is rank deficient if $p \leq \lfloor \frac{n+1}{2} \rfloor - 1 = \lfloor \frac{n-1}{2} \rfloor$, compare with (5). One can then recover the p poles α_i by rooting the polynomial associated with any vector in its kernel.

The algorithm to be used to recover the α_i 's, searching for eigenvectors and the roots of a polynomial, is now more complex and of a quite different nature than solving a linear program, but in this context the roots need not be real, simple and lying on a grid.

6. REMARKS ON THE UNIQUENESS OF THE SOLUTION

The result we established in Lemma 2 is valid for both normalized or un-normalized columns in A . In the last case, see (3), the first row of A is filled with ones and the first equation in $Ax = b$ is $\mathbf{1}^T x = b(1)$ where $b(1)$ denotes the first component of b . Since the criterion in (LP+) is precisely $\min \mathbf{1}^T x$, this means that all admissible points have the same cost $b(1)$. Since we proved uniqueness of the optimal solution, we actually established uniqueness of the admissible points.

Lemma 3: For A a (n,m) -Vandermonde matrix with distinct columns, the solution to $\{Ax = b, x \geq 0\}$ is unique if $b = Ax_o$ with $x_o \geq 0$ and

$$\|x_o\|_0 \leq \lfloor \frac{n-1}{2} \rfloor$$

where $\lfloor z \rfloor$ denotes the nearest integer that is $\leq z$. \square

This actually means that the cost function in (LP) is accessory, solving (LP) is just a mean to find the unique admissible point. Indeed one can reach the same conclusion without specifying any cost function by observing that the vector f we introduced in the proof of Lemma 2 verifies the following proposition:

Proposition: The solution x_o of $\{Ax = b, x \geq 0\}$ with $b = Ax_o = \bar{A}_o \bar{x}_o$ is unique if

$$\exists f \ni \bar{A}_o^T f = 0 \text{ and } a_j^T f > 0 \ \forall a_j \notin \bar{A}_o \quad (6)$$

with \bar{A}_o a full column rank matrix. \square

This result applies only for $\|x_o\|_0 < n$, since otherwise $f = 0$. This proposition is a consequence of the following lemma we will prove.

Lemma 4: The j -th component of all points in the set, assumed to be non-empty, $\{Ax = b, x \geq 0\}$ is zero if and only if $\exists f \ni b^T f = 0, A^T f \geq 0, a_j^T f > 0$. \square

Proof: To establish the necessity of the condition, one introduces the linear program

$$\min -e_j^T x \text{ subject to } Ax = b, x \geq 0$$

with e_j is the j -th column of the identity matrix. Its dual is

$$\max b^T d \text{ subject to } A^T d \leq -e_j.$$

With x_o the optimum of the primal, that exists by assumption and has cost zero, one can then associate d_o an optimum of the dual which satisfies $b^T d_o = -e_j^T x_o = 0$, $A^T d_o \leq 0$ and $a_j^T d_o \leq -1$. One then takes $f = -d_o$ to establish the result. The condition is sufficient since it implies $f^T Ax = f^T b = 0$. But $f^T Ax = 0$ with $f^T A \geq 0$ and $x \geq 0$ in turn implies that $e_j^T x = 0$ whenever $f^T a_j > 0$. \square

The proposition then follows from Lemma 4 and the assumption that \bar{A}_o is a full column rank matrix.

7. CONCLUSIONS

We have shown that results recently developed in the 'sparse representations' context can be used in realization theory. The application we proposed is quite simple but other applications can be developed. The same analysis performed in the presence of noise would correspond to the case where one tempts to simultaneously estimate the order and identify the coefficients of an AR process using estimates of the partial covariance sequence. The same could possibly be done for ARMA processes. The interest lies however more in fact that one substitutes solving a linear program to computing eigenvalues, eigenvectors, rooting polynomials, two completely different kind of operations.

A similar analysis can be performed when A is a real Fourier matrix with columns

$$\begin{aligned} s(\omega) &= \frac{1}{2}(v(e^{i\omega}) + v(e^{-i\omega})) \\ &= [1 \cos \omega \cos 2\omega \dots \cos k\omega \dots \cos(n-1)\omega]^T. \end{aligned}$$

indeed since $s(\omega)$ can be written as $Tv(\cos(\omega))$ with T a square lower-triangular invertible matrix built upon the coefficients of the Tchebychev polynomials, the two problem are equivalent. In this trigonometric context, the unicity result above has to be related to a trigonometric moment problem solved by Caratheodory [11].

One can also note that the results in Lemmas 1, 2 and 3 can also recovered using ideas from the theory of convex polytopes [12] or more precisely from the notion of k -neighbourly polytopes mainly developed by Gale [13] and applied to this context by Donoho [14].

8. REFERENCES

- [1] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24, 21, 227-234, April 1995.
- [2] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. on I.T.*, 47, 11, 2845-2862, Nov. 2001.
- [3] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. on I.T.* 49, 12, 3320-3325, Dec. 2003.
- [4] J.J. Fuchs. More on sparse representations in arbitrary bases. *IEEE Trans. on I.T.* 50, 6, 1341-1344, June 2004
- [5] J.A. Tropp, "Greed is good: Algorithmic Results for Sparse Approximations," *IEEE Trans. on I.T.*, 50, 10, 2231-2242, Oct. 2004.
- [6] D. G. Luenberger. Introduction to linear and nonlinear programming. *Addison Wesley*, 1973.
- [7] J.J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise *IEEE Trans. on I.T.* 51, 10, 3601-3608, Oct. 2005.
- [8] T. Strohmer and R. Heath, "Grassmanian frames with applications to coding and communications," *Appl. Comp. Harm. Analysis*, 14, 3, 257-275, May 2003.
- [9] A. Gilbert, S. Muthukrishnan and M.J. Strauss, "Approximation of functions over redundant dictionaries using coherence," *14th ACM-SIAM Symposium on Discrete Algorithms*, (SODA'03), Jan. 2003.
- [10] T. Kailath. Linear systems. *Prentice Hall*, 1980.
- [11] U. Grenander and G. Szego. Toeplitz forms and their applications. *Berkeley, Univ. Calif. Press*, 1958.
- [12] B. Grunbaum. Convex polytopes Graduate Texts in Mathematics, vol. 221, Springer NY, 2003.
- [13] D. Gale. "Neighboring vertices on a convex polyhedron" *Annals of Mathematical Studies*, 38, pp. 255-263, Princeton Univ. Press, N.J.? 1956.
- [14] D.L. Donoho and J. Tanner "Sparse Nonnegative solution of underdetermined linear equations by linear programming" Tech. Report 2005-6, Dept. Statistics, Stanford Univ., April 2005.