Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

On the use of sparse representations in the identification of line spectra.

Jean Jacques Fuchs *

* *Irisa-Université de Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France.(e-mail: fuchs@irisa.fr).*

Abstract: Sparse representations is a technique that consists in decomposing a signal into a small number of components, chosen from a user-designed over-complete set of vectors. While it is mostly used to obtain an approximate model of a signal or image, for compression or coding purposes, it can also be applied for identification, estimation or even detection purposes, when there exists a true exact sparse representation, which is then the object of interest. We consider the basic problem of the identification of real sinusoids in noise. While, in case of regular sampling the competition is formidable, for irregular sampling, it is far less exacting. The approach, we propose, applies to irregular samples without additional difficulty and attains performances close to the Cramer-Rao bound for quite reasonable computational costs.

1. INTRODUCTION

The estimation of the power spectral density of wide sense stationary random processes has received considerable attention (Porat [1994], Stoica [1997]). It has applications in many different fields. In this general domain, the very specific case, where the process can be modeled as the sum of real sinusoids in white noise, occupies a central position and has certainly caught more attention than any other. We consider this case and propose to estimate both the number of sinusoids, that are present and their characteristics. Even for this quite specific problem, there does not seem to exist a single approach working well in all situations. We thus further confine our attention to configurations where the number of observations is quite small (around 100) and the Signal to Noise Ratio (SNR) small also (below 10 dB per sinusoid). This is an often considered domain in the literature, that covers many situations in practice. It is also a domain where it is difficult to outperform the basic periodogram (Fourier transform), since the SNRs are quite low. Indeed, while so-called high-resolution methods have the ability to separate two sinusoids with closely spaced frequencies provided the SNRs are large enough, and while the periodogram is not a high resolution method, it is difficult to outperform it, at low SNRs. A large number of methods have been proposed, see (Stoica [1997]) and the references therein, and some of them do indeed separate sinusoids that the periodogram cannot resolve, for, say, SNR's above 0 dB.

The method we propose, has this capacity, but more importantly, it works with irregularly sampled data without any additional difficulty, and methods that can handle this type of data are extremely scarce. While the periodogram, the maximum likelihood and some model-fitting type approaches can be applied, in practice performance and feasibility will be an issue. Since the periodogram (the Fourier spectrum of the data) is the convolution of the true spectrum (a few lines) and the *spectral window* (the Fourier spectrum of the sampling instants) that can exhibit any pattern of secondary lobes, the potentiality of this approach is limited. As for the the maximum likelihood or model-fitting type approaches, in their usual implementation, their performance fully depends upon the quality of their initialization. The major difficulty for these methods is the availability of a good initial point and we are back to the initial problem. Not to mention, the necessary preliminary estimation of the number of sinusoids that are present, that is required by most techniques.

The method, we propose, can indeed be seen as a model-fitting approach, that performs simultaneously the estimation of the number of components and the identification of their characteristics. It furthermore requires no initialization point, since it relies upon the minimization of a convex function. It is also completely un-sensitive to the way the data have been sampled provided, of course, the sampling instants are known. We proposed it initially, in a different context, in (Fuchs [1997]) and (Fuchs [2001]) and it can now be seen as an application of the "sparse representations" techniques, a topic that has expanded trough many areas in signal processing in recent years (Donoho [2001], Gribonval [2003], Fuchs [2004]), and whose theoretical aspects concentrate more recently on the so-called compressed-sensing or compressed-sampling area (Candes [2006], Donoho [2006], Candes [2006b]).

In the control community, besides applications in identification (Zhang [1999]), "sparse representations" techniques can also be of some use in realization theory (Fuchs [2006]) or failure diagnosis (Merrill [1973]). While in most cases, sparse representations techniques are used to decompose a signal into a small number of components chosen from user-defined redundant set of vectors, it can also be applied, for detection or estimation purposes, when there exists a true exact sparse representation. It is this last point of view, that we adopt here and that is mostly of interest to the control community. But, it can, quite generally, be adapted as soon as an observation (a set of data or a set of estimated covariances) can be represented as the sum of a small number of vectors belonging to a known family (sinusoids, shifted replicas of a same signal, failure signatures, ..) in additive noise (measurement noise, estimation errors).

In section 2, we detail the identification problem we shall consider. We specify the criterion to be minimized in Section 3 and discuss the implementation issues in Section 4. In Section 5, we present some simulation results and after some concluding remarks, we indicate in the Appendix how to build a quite efficient minimization algorithm.

2. THE MODEL

We consider the following noise corrupted sinusoidal signal:

$$y(t_k) = \sum_{j=1}^{P} A_j \cos(2\pi f_j t_k + \varphi_j) + e(t_k) \tag{1}$$

where e_{t_k} is zero mean white Gaussian noise with variance σ^2 and the initial phases φ_j are assumed independent random variables uniformly distributed in $[0,\ 2\pi[$. We address the problem of the estimation of the number P of sinusoids together with the identification the positive amplitudes A_j, the initial phases $\varphi_j \in [0,\ 2\pi[$, and the frequencies $f_j \in [0,\ .5[$, normalized with respect to a standard unity sampling period, from a set of data $\{y(t_k)\}$ of length n. We will assume that the irregular sample instants t_k satisfy $0 \le t_1 < t_2 < .. < t_n \le n-1$, i.e., that in the average, the irregular samples satisfy the Nyquist rate (Yen [1956]).

It will appear clearly from the approach we use, that irregular sampling induces strictly no additional complexity in the implementation. The difference may be a slightly increased computation time, since the conditioning of the convex function to be minimized may be affected, and a change in the statistical performances. Indeed, under the Gaussian noise assumption, the proposed method may well attain the Cramer-Rao bounds and these will depend upon the sampling scheme. These excellent performances of the proposed approach are difficult to establish theoretically, but can be observed on the simulation results.

3. THE CRITERION

Using the n data points $\{y(t_k)\}$, we build a n-dimensional column vector $\mathbf{y}$ and define similarly vectors $\mathbf{s}(f_j,\ \varphi_j)$ and a noise vector $\mathbf{e}$ which are such that (1) can be rewritten

$$\mathbf{y} = \sum_{j=1}^{P} A_j \mathbf{s}\ (f_j, \varphi_j) + \mathbf{e}. \tag{2}$$

The sparse representation technique, we propose, then simply amounts to search a sparse (the sparsest) decomposition of $\mathbf{y}$ as a linear combination of column vectors $\mathbf{a}_{p,q}$, of the form:

$$\mathbf{a}_{p,q} = \mathbf{s}\ (\frac{p-1}{2n_f},\ \frac{2\pi(q-1)}{n_\varphi}), \tag{3}$$

with $p \in (1, n_f)$ and $q \in (1, n_\varphi)$. One aims to reconstruct (the deterministic part of) $\mathbf{y}$ as the sum of sinusoids whose frequencies and initials phases are on a 2-dimensional grid. There are indeed an infinite number of ways to achieve it but one expects the sparsest such decomposition to be close to the true one. One seeks the vector x of dimension $n_f \times n_\varphi$ having the fewest number of non-zero components, such that $\mathbf{y} \simeq Ax$ where A is the matrix having the $\mathbf{a}_{p,q}$'s as columns.

In fact, since observation noise $\mathbf{e}$ is present in (2), we will not try to reconstruct $\mathbf{y}$ exactly. If the discretization steps in frequency and initial phase are small, or, equivalently, if n_f and n_φ are large, P vectors $\mathbf{a}_{p,q}$ may then be sufficient, since the allowed reconstruction error may also take care of the errors between $\mathbf{s}(f_j, \varphi_j)$ in (2) and the nearest $\mathbf{a}_{p,q}$-vector present in the matrix A. Since the noise is assumed to be Gaussian, we use the Euclidean norm to measure the reconstruction error $Ax - \mathbf{y}$ and propose to solve

$$\min_x \|x\|_1, \quad \text{s.t.} \quad \|Ax - \mathbf{y}\|_2^2 \le \rho \tag{4}$$

where $\|x\|_1 = \sum |x_j|$ denotes the ℓ_1-norm and $\|x\|_2^2 = \sum x_j^2$ is the square of the Euclidean norm. Whatever the value of the allowed discrepancy ρ, there are an infinite number of admissible points in (4), and one replaces the true exhaustive search for the sparsest such point by the easy minimization of its ℓ_1-norm. The quasi-equivalence of these two point of views is considered in (Donoho [2001], Gribonval [2003]). Now (4) is equivalent to

$$\min_x \frac{1}{2}\|Ax - \mathbf{y}\|_2^2 + h\|x\|_1, \quad h > 0 \tag{5}$$

where h has to be tuned by the user, just like ρ in (4). This is the criterion that is mostly used in the overwhelming literature on “sparse representations’,’ and already considered, for instance, in (Fuchs [1999]). This criterion is convex, as the sum of convex functions, it has, thus, a unique minimum that is generically attained at a unique point, when $\mathbf{y}$ is a noisy data vector. It can be transformed into a quadratic program, but efficient algorithm specifically dedicated to solve (5) have been developed (Osborne [2000], Efron [2004], Maria [2006]). These algorithms converge in a finite number of steps and this number is about twice the number of non-zero components in the optimum of (5). Since in our context, this number is of the order of P, the number of sinusoids, this means that one can easily consider situation where A has a few thousands of columns. We detail in the Appendix, both for completeness and since, to our knowledge, it has never been published, the development of an algorithm of this type, that solves (5) under the additional constraint $x \ge 0$ which can be enforced in our application.

4. IMPLEMENTATION ISSUES

4.1 Recovery of the estimates

Let us indicate how we deduce the estimates of the parameters in (1) from the optimum of (5).

If x_0 denotes the optimum, we first transform this column vector into a matrix, say, X_0 of dimension (n_f, n_φ). In case, a given sinusoid in (1, gives rise to several non-zero components in x_0, one expects these to form a (compact) cluster in the 2-dimensional matrix X_0. From such a cluster, we will deduce the estimates of the frequency and the initial phase of the given sinusoid by linear interpolation. Since the rows in X_0 are associated with the frequencies, in case there are several columns in the considered cluster, we sum over the columns to get the weights to be used in the linear interpolation to yield the frequency estimate. We proceed similarly to get the initial phase estimate. The number of (significant) clusters in X_0 then determines the estimate of P the number of sinusoids present in the observations. In general and in the simulations given below, a sinusoid gives rise to between 1 and 3 non-zero components and if some additional components are present they are always isolated and of very small magnitude.

4.2 Choice of the discretization steps

With the notations introduced in Section 3 and below relation (3), the discretization step is $\delta_f = 1/(2n_f)$ in frequency and $\delta_\varphi = 2\pi/n_\varphi$ in initial phase. We choose these steps so that they do not prevent the approach to attain its best performances, i.e., the Cramer-Rao bounds, For equispaced sampling, one has a fairly good idea of these bounds (in resolution), and for the SNRs we consider and with n denoting the number of data points, we propose to take $n_f = 4n$, in order to always have potentially a few non-zero components even between closely spaced sinusoids. As far as δ_ϕ is concerned, we will in general take $\delta_\varphi = \pi/6$ or $\pi/8$ and sample the domain $[0, 2\pi[$ if we minimize (A.1) with positive weights in x, or sample the domain $[0, \pi[$ if we minimize (A.1) with arbitrary weights in x.

For $n = 60$ regularly spaced sample points, this leads to a A-matrix with 60 rows and about 2000 or 4000 columns, respectively. One can actually prove that the linear interpolation procedure allows to gain an order of magnitude in the precision, i.e., it somehow transform the discretization step δ into δ^2. For irregularly sampled data, the performances are no longer uniform and the above rule for the choice of δ_f is no longer justified, we nevertheless propose to keep it.

4.3 Choice of the threshold h

This is, of course, an important parameter, since, as explained in the Appendix, the number of non-zero components in the optimum of the criterion directly depends upon h. Roughly speaking, their number increases (but never exceeds n), as h decreases.

To get a feeling of the meaning of the threshold h, we introduce, the Lagrangian dual of (5) (Fuchs [2001])

$$\min_x \|Ax\|_2^2 \text{ s.t. } \|A^T(Ax - \mathbf{y})\|_\infty \leq h. \qquad (6)$$

This problem is strictly equivalent to (5) and its constraint tells us that, at the optimum, the correlation of the reconstruction error $Ax - \mathbf{y}$ with any column in A^T (column of A) is smaller than h in absolute value. In the ideal case where at the optimum x_0, Ax_0 represents exactly the contribution of the P sinusoids to $\mathbf{y}$, the reconstruction error $Ax_0 - \mathbf{y} = \mathbf{e}$ and the constraints reads $\|A^T\mathbf{e}\|_\infty \leq h$. From this observation, one first deduces that it is essential to normalize the columns in A to one, say, in Euclidean norm to guarantee they are all given the same importance. It follows then that the components in $A^T\mathbf{e}$ are (dependent) Gaussian random variables with mean zero and variance σ^2, and, to make this ideal model admissible, we fix h at a value that guarantees that the probability that the maximum of these $n_f \times n_\varphi$ random variables be larger than h is close to zero. Following (Leadbetter [1983], Fuchs [2001]) we suggest to take $h = \sigma\sqrt{\log 2n_f n_\varphi}$.

4.4 The complete algorithm

We summarize the algorithm we propose and for which we present simulation results in the next section. Given n regularly or irregularly sampled data points following model (1), we form a n-dimensional vector $\mathbf{y}$, we take $n_f = 4n$ and $n_\varphi = 12$ and form the $n_f \times n_\varphi$ vectors $a_{p,q}$ defined in (3), we normalize these vectors and build the matrix A having them as column vectors. We then solve

$$\min_{x \geq 0} \frac{1}{2}\|Ax - \mathbf{y}\|_2^2 + h\|x\|_1, \quad h > 0$$

using the algorithm sketched in the Appendix, with $h = \sigma\sqrt{\log 2n_f n_\varphi}$. We eventually deduce the estimates of A_j, f_j, φ_j from the optimum, as explained in section 4.1.

To simplify the exposition, we assume the variance σ^2 to be known and do not detail how the detection of the number P of sinusoids could be performed. We will however indicate, for the 2 sets of simulations, the value of the maximal non-zero component remaining in the optimum once the components associated with the sinusoids that are present, have been removed.

5. SIMULATION RESULTS

We consider two closely spaced sinusoids.

5.1 The regular samples case

We take $n = 60$ samples that are uniformly sampled with a period equal to one, We consider a frequency separation of $1/2n$ and identical initial phases, to fix ideas. The signal to noise ratio (SNR) defined to be equal to $A_j^2/2\sigma^2$ is taken equal to 10. We assume to know σ^2 which we take equal to one. We apply the algorithm described in Section 4.4 with $n_f = 4n$ and $n_\varphi = 16$, this leads to a matrix A of dimension (60, 3840).

Taking identical initial phases is a favorable situations, but the frequency separation $1/2n$ is quite small and only so-called high resolution methods can separate the two sinusoids. We present in Table 1, the mean and the variance of the estimates obtained over 10000 independent realizations as well as the Cramer Rao bounds (CRB). The proposed approach separates the two sinusoids and almost achieves the CRB for the frequency estimates. The amplitude estimates are biased but this is a side effect of the criterion and can be corrected.To further improve the performance, one could initialize any optimization algorithm maximizing the likelihood, it would converge to the maximum likelihood estimates in a few steps. To complete the picture, let us indicate that, the amplitude of the maximal additional non-zero component (if any) of the optimal x was 0.72 in the average over the 10000 realizations. This means that in case one does not known the number of sinusoids that are present, then, if our approach detects an additional one, its amplitude is generally quite small (seven times smaller that the two true ones) and it may potentially be declared to weak and discarded.

Table 1: Estimates of the means and variances averaged over 10000 independent realizations

	Mean	Variance	CRB
$A_1 = 4.47$	3.93	0.376	0.165
$A_1 = 4.47$	3.88	0.358	0.167
$f_1 = 0.2502$	0.2498	$1.31.10^{-6}$	$0.82.10^{-6}$
$f_2 = 0.2586$	0.2591	$1.26.10^{-6}$	$0.84.10^{-6}$
$\varphi_1 = 1.57$	1.89	0.010	0.044
$\varphi_1 = 1.57$	1.24	0.011	0.045

Two equipowered sinusoids in white noise. The frequency separation is 1/2 of the Raylegh limit. SNR=10 dB, n=60 data points, $\Delta f = 1/120$, identical initial phases.

5.2 The irregular samples case

We now turn to irregular sampling, the modification to be made to the proposed is quite simple, one applies the same irregular sampling scheme when building the columns of the A matrix. We consider a similar scenario and keep the same characteristics for the two sinusoids except for the frequency separation taken equal to $2/3n$, i.e., slightly larger but still below the Raylegh limit, $1/n$, valid for equispaced samples. The same irregular sampling scheme is kept for all the realizations. In place of the regular samples between 1 and 60, we consider three clusters of 20 points each, randomly distributed around the values 10, 30 and 50. A typical realization of the irregularly sampled data points is presented in Figure 1. In Figure 2, we present the spectral window associated with the specific sample-point-set we used in our simulations. In case of regular samples this spectral window is the so-called Fejer kernel (Porat [1994], Stoica [1997]). In Figure 3, we present the interesting part of the periodogram obtained for a typical data realization together with the locations of the two true spectral lines and those of the average estimates we get. Looking at the periodogram in Figure 3, it is indeed difficult to guess that there are two spectral lines, slightly shifted to the right, in the central main lobe with no other spectral lines elsewhere.

The frequency estimates, we get, present a slight outward bias but the the mean square error is quite close to the CR bounds. To complete the picture, let us indicate that, the amplitudes of the maximal additional non-zero component (if any) of the optimal x was 0.64 in the average over the 10000 realizations.

6. CONCLUDING REMARKS

We have presented an algorithm for estimating the parameters of sinusoidal signals in noise. It can also be applied to others identification problems. Its performance are quite good and close to the Cramer Rao bounds. It has at least two advantages over other approaches, it requires no initialization point and can handle irregularly samples

Table 2: Estimates averaged over 10000 independent realizations in the irregular sampling case, see Figure 1.

	Mean	Variance	CRB
$A_1 = 4.47$	3.36	0.069	0.197
$A_1 = 4.47$	3.49	0.081	0.177
$f_1 = 0.2502$	0.2485	$0.87.10^{-6}$	$0.86.10^{-6}$
$f_2 = 0.2614$	0.2627	$0.92.10^{-6}$	$1.27.10^{-6}$
$\varphi_1 = 1.57$	2.01	0.007	0.059
$\varphi_1 = 1.57$	1.13	0.011	0.071

Two equipowered sinusoids in white noise. The frequency separation is 2/3 of the Raylegh limit. SNR=10 dB, n=60 data points, $\Delta f = 1/90$, identical initial phases.

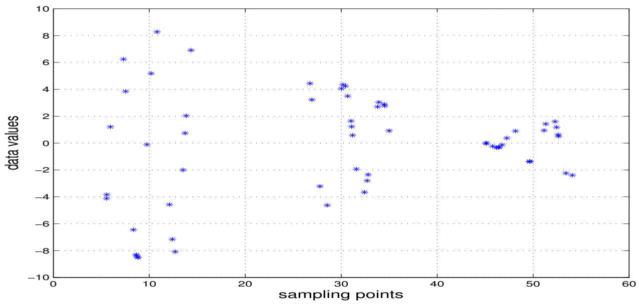

Fig. 1. The irregularly sampled sum of two sinusoids, there are 3 clusters of 20 samples each, around the samplings instants 10, 30 and 50.

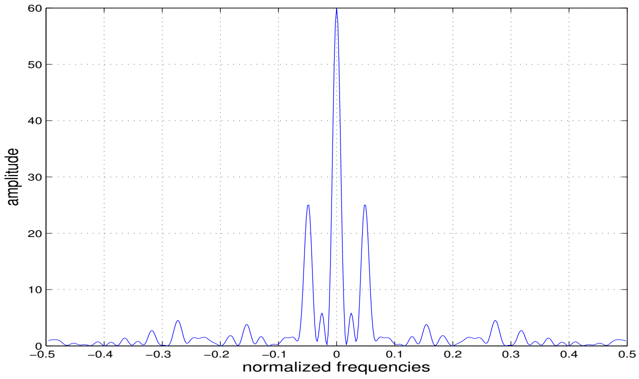

Fig. 2. The spectral window associated with the irregular samples.

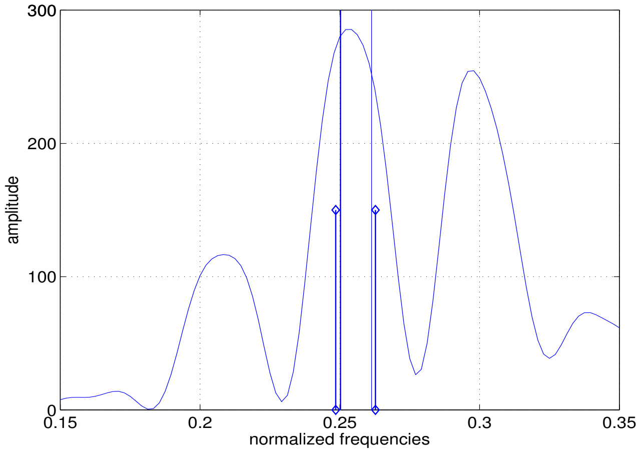

Fig. 3. Zoom on the periodogram of a typical data set, the diamonds locate the estimates we get, they are close to the true spectral (vertical) lines.

data without any additional difficulty. Its major drawback is probably its computational cost but we present in the Appendix, the ingredients that allow to build a highly efficient algorithm able to solve the required optimization problem in a few tenth of seconds on a laptop, though this problem has several thousands of unknowns.

REFERENCES

B. Porat. Digital processing of random signals. *Prentice Hall.* N.J.,1994.

P Stoica and R. Moses. Introduction to spectral analysis. *Prentice Hall.* N.J., 1997.

J.J. Fuchs. Extension of the Pisarenko method to sparse linear arrays *IEEE-T-SP*, 45: 2413–2421, Oct. 1997.

J.J. Fuchs. On the application of the global matched filter to DOA estimation with uniform circular arrays. *IEEE-T-SP*, vol. 49, p. 702–709, avr. 2001.]

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. on I.T.*, 47, 11, 2845-2862.

R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. on I.T.* 49, 12, 3320-3325, Dec. 2003.

J.J. Fuchs. More on sparse representations in arbitrary bases. *IEEE Trans. on I.T.* 50, 6, 1341-1344, June 2004, also in *13th IFAC SYSID*, 1357-1362, Rotterdam, 2003.

E. Candes, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information." *IEEE T. on I.T.*, 52, 489-509, 2006.

D. Donoho, "Compressed sensing." *IEEE Trans. on I.T.*, 52, 1289-1306, 2006.

E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal coding strategies ?" *IEEE Trans. on I.T.*, 52, 5406-5425, 2006.

Q. Zhang and J.J. Fuchs, "Building neural networks through linear programming." *14th IFAC World Conf.*, K, 137-142, 1999, Beijing.

J.J. Fuchs. Sparse representations and realization theory. *MTNS* Kyoto, 2006.

H.M. Merrill, "Failure diagnosis using quadratic programming" *IEEE Trans. on Reliability*, 22, 207-213, 1973.

J.L. Yen. On nonuniform sampling of bandwidth-limited signals. *IRE -Trans. Circuit Theory* 3, 251–257, 1956.

J.J. Fuchs. Multi-path time-delay estimation. *IEEE-T-SP*, vol. 47, p. 237–243, Jan. 1999.

M. Osborne, B. Presnell and B. Turlach. "A new approach to variable selection in least squares problems" *IMA J. of Numerical Analysis*, 20, 3,389–403, 2000.

B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, "Least angle regression," *Annals of Statistics*, 32, pp. 407–499, Apr. 2004.

S. Maria and J.J. Fuchs, "Application of the Global Matched Filter to STAP data: an efficient algorithmic approach." In *Proceedings ICASSP*, Toulouse, may 2006.

M.R. Leadbetter, G. Lindgren and H. Rootzen. Extremes and related properties of random processes and sequences. *Springer Verlag*, 1983.

D.G. Luenberger. Introduction to linear and non-linear programming. *Addison Wesley*, 1974.

Appendix A. A SKETCH OF THE ALGORITHM

We explain how to build an iterative algorithm that solves

$$\min_{x\geq 0} \ \frac{1}{2}\|Ax-y\|^2 + h\|x\|_1, \quad h>0 \tag{A.1}$$

with A is $n\times m$ matrix of rank n, b and x are real vectors of adequate dimension and h is a positive real. Note that since $x\geq 0$, one has $\|x\|_1 = \sum x_k = \mathbf{1}^T x$.

Since the criterion is convex, it is well known that (Luenberger [1974])

Lemma 1. x is a minimum of (A.1) if and only if

$$\exists \ \ \mu \geq 0 \ \ \ni A^T(Ax-y) + h\mathbf{1} - \mu = 0, \tag{A.2}$$

with $x^T\mu = 0$. □

We denote x the optimum of (A.1) that thus also satisfies (A.2) and split it into its strictly positive components we denote $\bar{x}$, and its zero components we denote $\bar{\bar{x}}$. We partition, similarly and accordingly, A and μ. From $x^T\mu = 0$, it then follows that $\bar{\mu} = 0$ and $\bar{\bar{\mu}}\geq 0$. Using these notations, we decompose (A.2) into

$$\bar{A}^T(y-\bar{A}\bar{x}) = h\mathbf{1} \ \text{ and } \ \bar{\bar{A}}^T(y-\bar{A}\bar{x}) = h\mathbf{1} - \bar{\bar{\mu}}$$

One can now draw $\bar{x}$ from the first relation and replace it in the second to get

$$\bar{x} = \bar{A}^+ y - h(\bar{A}^T\bar{A})^{-1}\mathbf{1} \tag{A.3}$$

$$\bar{\bar{\mu}} \ = -\bar{\bar{A}}^T y^\perp - h\,\bar{\bar{A}}^T \ (\bar{A}^{+T}\mathbf{1} - \bar{\bar{\mathbf{1}}}) \tag{A.4}$$

with $\bar{A}^+ = (\bar{A}^T\bar{A})^{-1}\bar{A}^*$, where we assume the inverse to exist, and $y^\perp = (I - \bar{A}\bar{A}^+)y$.

One can now observe that, if one knows the optimum x for a given h, then using (A.3) and (A.4), one can extend it to a interval around the current h, the interval for which both $\bar{x}$ and $\bar{\bar{\mu}}$ remain ≥ 0. But it is also possible extend the optimum to the neighboring intervals, and, thus, to all $h>0$. It only remains to get a starting value, i.e. the optimum for a given value of h, and this is easy, since for h large, the optimum is at zero. The idea is the same as the one used in (Osborne [2000],Efron [2004],Maria [2006]).

More precisely, from (A.2) it follows that for $h\geq h_0 = \max(A^Ty)$ the optimum is at $x=0$. For h in an interval $[h_1, \ h_0[$ with $h_1 < h_0$ yet to be defined, the optimum of (A.1) has then just one nonzero component with index $j_1 = \arg\max_k a_k^T y$. For h in this interval, $\bar{A} = [a_{j_1}]$ and $\bar{\bar{A}}$ contains all the remaining columns of A. From (A.3), it follows that $\bar{x} = x_{j_1} = (a_{j_1}^T y - h)/(a_{j_1}^T a_{j_1})$. As h decreases within this interval, and, more generally, as h decreases from, say, h_k, the values in $\bar{x}(h)$ (A.3) and $\bar{\bar{\mu}}\ (h)$ (A.4) vary linearly and the next boundary value h_{k+1} is the value for which either a component in $\bar{x}(h)$ or $\bar{\bar{\mu}}\ (h)$ becomes zero (first), one then modifies the partitions, moving a column from $\bar{A}$ to $\bar{\bar{A}}$ or vice versa, and proceeds.

In summary to solve (A.1), though one is only interested in the optimum for a given value of h, it happens to be cheaper to solve the problem for decreasing h, i.e., to decompose the positive real axis into intervals $]h_{k+1}, \ h_k]$ within which the number of nonzero components of the optimum remains constant and to stop the procedure when the h of interest is within the current interval. The algorithm one gets is highly efficient and especially so when the number of non-zero components in the sought optimum is small. Since a component of x, which is non-zero for a given h, may become zero again later, the number of steps (intervals) required to solve (A.1), is greater than the number of non-zero components in the optimal x and generally about twice this number.