

## SPARSITY AND STRUCTURE IN AUDIO SIGNAL THROUGH MIXED NORMS

MATTHIEU KOWALSKI

Sparse and structured signal expansions on dictionaries (i.e. a collection of elementary waveforms, called atoms) can be obtained through explicit modeling in the coefficient domain. For music signals, the most popular choice for dictionaries is the time-frequency representations. However, Heisenberg’s uncertainty principle denies the possibility to have a good frequency resolution and a good time resolution at the same time. In order to represent both time structure such *transients* (for example, percusif sounds) and frequency structures such *tonals* (for example, musical notes played by a violin), one can choose an over-complete dictionary contained the two kinds of atoms. Then, on can seek relevant waveforms using a sparse representation of the signal.

However, using only sparsity is an over-simplification of reality. If, in a well chosen dictionary, a signal has some “large” coefficients, some structures clearly appear. Fig. 1 gives two time-frequency representations of a glockenspiel signal. At least two kinds of structures can be observed:

- (1) A structure in layers: depending of the chosen dictionary, the signal present different properties. With a short time analysis window, one can observed transients which corresponds to the hit on the bell. With a long time window analysis, one can observed tonal structures corresponding to the partial harmonics.
- (2) Inside each layer, some structures appear. Transients, well localized in time, have a large frequency coverage and tonals present harmonic structures.

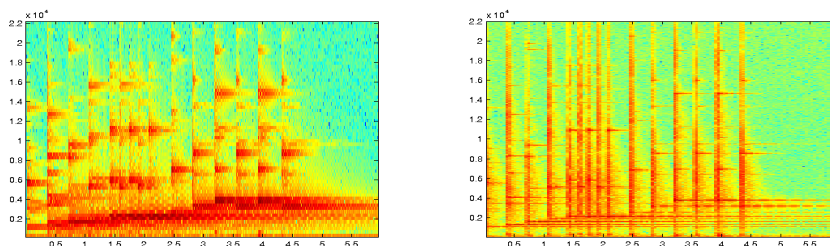


FIGURE 1. Time frequency MDCT coefficients of a glockenspiel signal. Left: short analysis. Right: Long analysis window.

A classical formulation to deal with sparsity is the LASSO [1] or Basis Pursuit Denoising (BPDN) [2]. Given a (noisy) observation of a signal  $s$ :  $x = s + b$ , one looks for a estimate of  $s$  supposed to admit a sparse decomposition inside a (time-frequency) dictionary. We denote  $\Phi$  the matrix which columns represent atoms of the dictionary, and  $\underline{s}$  synthesis coefficients of  $s$  such that  $s = \Phi \underline{s}$ . The LASSO or BPDN estimate of  $\underline{s}$  is given by

$$\hat{\underline{s}} = \underset{\underline{s}}{\operatorname{argmin}} \|x - \Phi \underline{s}\|_2^2 + \lambda \|\underline{s}\|_1$$

where  $\lambda \in \mathbb{R}_+$  is a parameter which control sparsity of the solution. An estimate  $\hat{s}$  of  $s$  is then obtained by  $\hat{s} = \Phi \hat{\underline{s}}$ .

A very strong hypothesis behind the  $\ell_1$  regularization is that all the synthesis coefficients  $\underline{s}$  are supposed to be i.i.d.. In order to introduce some structures in this approach, we propose to replace the  $\ell_1$  norm by a mixed norm [3,4]. Suppose  $\underline{s}$  be labelled by a double index  $(i, j)$  (for example, a time-frequency index). We call mixed norm of  $\underline{s}$  a  $\ell_{p,q}$  ( $p, q \geq 1$ ) norm defined by

$$\|\underline{s}\|_{p,q} = \left( \sum_i \left( \sum_j |s_{i,j}|^p \right)^{q/p} \right)^{1/q} .$$

Then, an estimate of  $\underline{s}$  can be obtained, solving the following problem

$$(1) \quad \hat{\underline{s}} = \underset{\underline{s}}{\operatorname{argmin}} \|x - \Phi \underline{s}\|_2^2 + \lambda \|\underline{s}\|_{p,q}^q .$$

Hence, one can favor sparsity accros the groups (with  $q$  close to 1) or inside each group (with  $p$  close to 1). With  $q = 1$  and  $p = 2$ , on recognize the Group-LASSO (also called joint sparsity, or multiple measurement vector). With  $p = 1$  and  $q = 2$ , we called the estimator the Elitist-LASSO.

If mixed norm can deal with some kind of structure, we did not deal with the structure in layers. For this aim, suppose that the signal  $s$  admit a decomposition like

$$s = s_1 + s_2 ,$$

where  $s_1$  can correspond to the transient layer, and  $s_2$  to the tonal one. We then construct a dictionary as an union of two dictionaries, each adapted for each layer. We denote by  $\Phi_1$  and  $\Phi_2$  the corresponding matrices. Regression problem 1 can be rewrite to deal with a layer decomposition of the signal, with a functional like

$$(2) \quad (\hat{\underline{s}}_1, \hat{\underline{s}}_2) = \underset{\underline{s}_1, \underline{s}_2}{\operatorname{argmin}} \|x - \Phi_1 \underline{s}_1 - \Phi_2 \underline{s}_2\|_2^2 + \lambda_1 \|\underline{s}_1\|_{p_1, q_1}^{q_1} + \lambda_2 \|\underline{s}_2\|_{p_2, q_2}^{q_2} .$$

One can then obtain an estimation of each layer:  $\hat{s}_1 = \Phi_1 \hat{\underline{s}}_1$  and  $\hat{s}_2 = \Phi_2 \hat{\underline{s}}_2$ . Of course, an estimation of  $s$  is given by  $\hat{s}_1 + \hat{s}_2$ .

All the functional can be optimized using iterative thresholding/shrinkage algorithms, after computing the corresponding shrinkage operator of the mixed norms [3]. Several applications can be formulate using mixed norm and/or a decomposition in layer: in particular, the author applied these techniques to audio signals for denoising tasks [3,4] and source separation of an under-determined convolutive mixture [5].

## REFERENCES

- [1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Serie B*, vol. 58, no. 1, pp. 267–288, 1996.
- [2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [3] M. Kowalski, "Sparse regression using mixed norms," 2008, submitted to *Applied and Computational Harmonic Analysis*.
- [4] M. Kowalski and B. Torr sani, "Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients," *Signal, Image and Video Processing*, 2008, doi:10.1007/s11760-008-0076-1.
- [5] M. Kowalski, E. Vicent, and R. Gribonval, "Under-determined source separation via mixed-norm regularized minimization," *Proceedings of the European Signal Processing Conference*, Aug. 2008.