Blind Source Separation of MALDI spectra from noise by sparse representation in a composite of linear bases

Luis Mancera[1], Luis González-Jaime[1], Matt Openshaw[2]

1 Clover Bioanalytical Software, Spain; 2 Shimadzu, Manchester, UK

Overview

- **We apply a Blind Source Separation technique to differentiate between signal and noise on MALDI spectra**.
- We look for efficient representations in a redundant composite of two transforms with different statistical properties, one adapted to represent the signal, the other the noise.
- Results show a clean separation between signal and noise with joint perfect reconstruction of the original signal

Sparse approximation on a composite of bases

- MALDI spectra is typically affected by several sources of noise: chemical background, electrical noise, peak broadening, contaminants...
- The signal of interest usually has statistical properties that differ from those of most of the sources of noise.
- The sparse approximation problem is defined as minimising the error of approximating the data as a linear combination of a limited number of functions taken from an overcomplete base (one with more coefficients than elements in the original signal)

We use a blind-source separation approach to differentiate between signal and noise by looking for sparse representations in a composite domain constructed from several bases, each one adapted to represent efficiently one of those sources.

- We implement the solution to the associated sparse approximation problem by using an iterated algorithm based on **Iterative Shrinkage Methods**, which iterates between
  - A thresholding operation in the transformed domain to enforce sparseness
  - A projection back to the transformed subspace of perfect reconstruction of the original signal

- By doing so, we hope to **capture each feature of the raw signal with the corresponding base** that is more adapted to its statistical properties.

## Experimental Methods

Sample preparation and acquisition:

- The proposed algorithm has been tested on MALDI-MSI datasets of mouse brain tissue sections. Serial mouse brain sections at 12 µm thickness were mounted on ITO glass slides for MALDI.
- Tissue sections were sprayed using the SunCollect (Sunchrom) for trypsin and matrices deposition. α-Cyano-4-hydroxycinnamic acid matrix was used for in-situ digested protein imaging respectively.
- MALDI images were acquired using a MALDI TOF-TOF instrument (MALDI-7090, Shimadzu, UK) in linear or reflectron mode for protein imaging and in-situ digest imaging samples, respectively at 30 µm resolution.

Processing:

- Raw data were loaded in IonView software (Shimadzu, UK) and TIC and Basepeak spectra were exported to ASCII format for further processing in Matlab.

Results & discussion

- Representations used:
  - 1D Dual-Tree Complex Wavelet transform
    - Represented by matrix **W**
    - **Efficiently represents signal**
  - 1D Discrete Fourier Transform
    - Represented by matrix **F**
    - **Efficiently represents noise**

- Algorithm
  - **x**; input raw spectrum
  - **A**: composite orthogonal transform
    - Column-wise concatenation of α**W** and β**F**, with α, β two scalars
  - $a^{(0)} = $ **Ax**, $i = 0$, $\theta^{(0)} = \max(abs(a^{(0)}))$, γ in $(0,1)$
  - Until convergence repeat:
    - Form $b^{(i)}$ by thresholding $a^{(i)}$ with threshold θ
    - Form $x_s^{(i+1)}$ and $x_n^{(i+1)}$ by transforming $b^{(i)}$ to the original signal domain of **W** and **F** respectively
    - Form $a^{(i+1)}$ by projecting $b^{(i)}$ to the subspace of perfect reconstruction of **x**
    - $\theta^{(i+1)} = \gamma\theta^{(i)}$
  - Use $x_s^{(end)}$ as estimate of noise-free signal
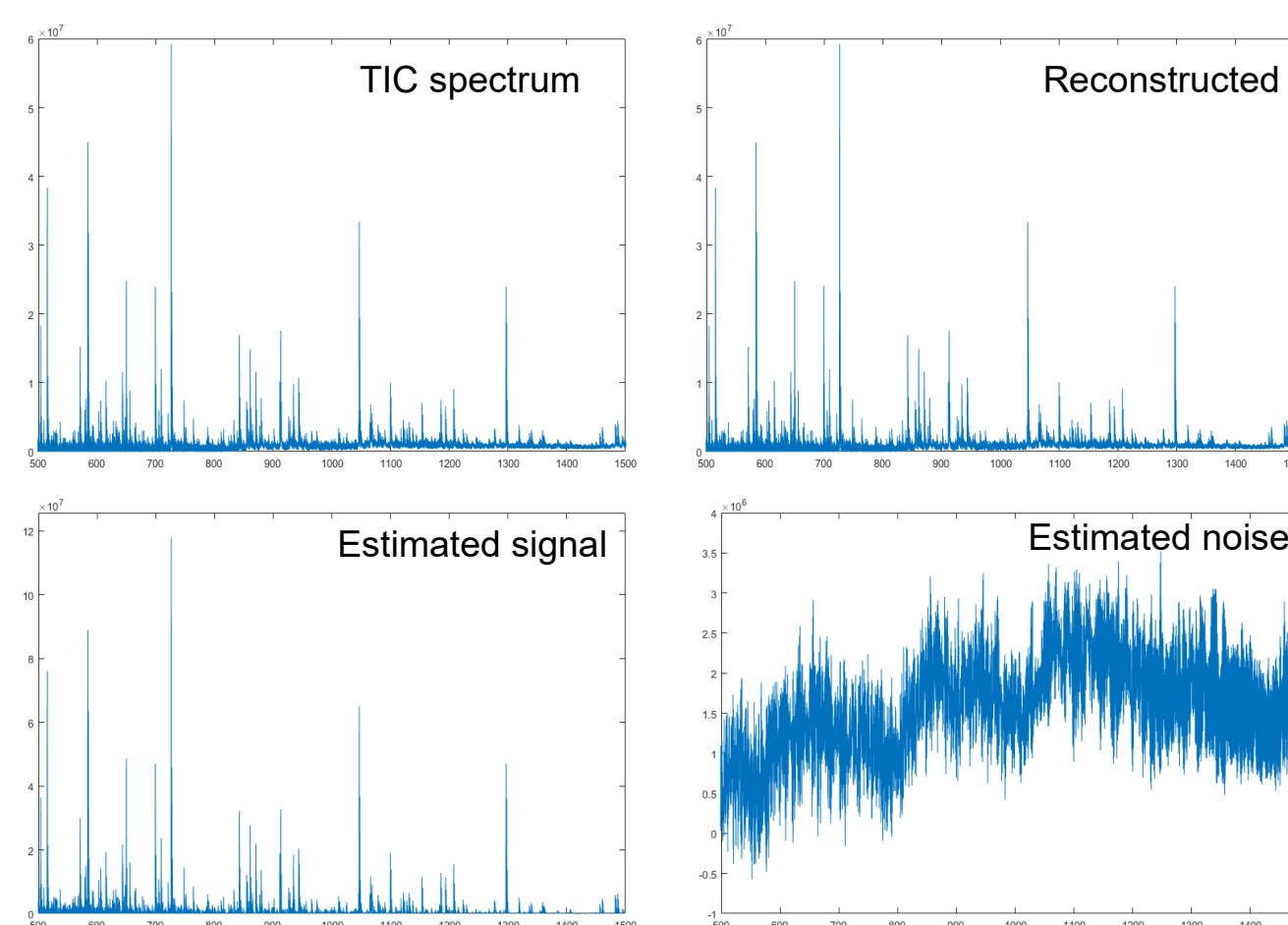
Figure 1: Noise separation result using a TIC spectrum from a mouse brain protein digest image. Mass range is showed from 500 to 1500 Da for feature visibility. Scalar parameters were set at $\alpha = 0.4$, $\beta = 0.6$, $\gamma = 0.9$. From top to bottom and left to right: original TIC spectrum, reconstructed spectrum from final sparse representation, estimated noise-free signal component, estimated noise component.
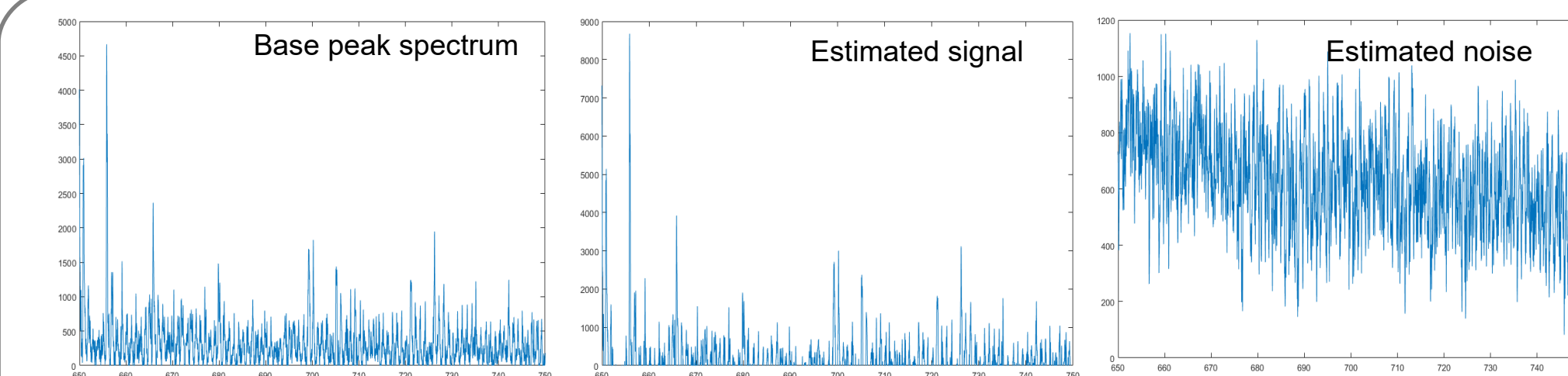
Figure 2: Same settings as in Figure 1 but applied to the Base peak spectrum. Mass range is showed from 650 to 750 Da for feature visibility.

Additional result: Data reduction

- Additionaly, this technique helps to reduce the disk size required to store spectra.
- This is thanks to the maximization of sparseness in the resulting representation in the transformed domain
- Compression results for the examples shown in this poster:

Example	Original data size	Compressed data size	Estimated signal data size	% of data reduction
TIC spectrum	2.03 MBytes	0.15 MBytes	0.06 MBytes	97.07%
Basepeak spe.	2.03 MBytes	0.70 MBytes	0.17 MBytes	91.53%

Conclusions

- This seems to be a promising approach to separate signal from noise in MALDI spectra. Noise-sources captured include random noise and baseline
- It is an iterative algorithm but it has the potential to be highly accelerated.
- It provides an estimation of the noise in the same range as the signal, which can be used to estimate noise power, calculate signal to noise ratios, etc.
- As side result, it provides the potential for highly compressing the data.

References

[1] J. Bobin, J-L Starck, J. Fadili, Y. Moudden. IEEE Trans. on Image Proc. (2007), 16: 2662-74

[2] A.N. Krutchinsky, B.T. Chait. J. Am. Soc. Mass Spectrom. (2002) 13(2):129-34.