SOURCE EXTRACTION USING SPARSE REPRESENTATION

Michael Zibulevsky *Yehoshua Y. Zeevi*

Department of Electrical Engineering
Technion, Haifa 32000, Israel
mzib@ee.technion.ac.il zeevi@ee.technion.ac.il

ABSTRACT

It was discovered recently that sparse decomposition by signal dictionaries results in dramatic improvement the qualities of blind source separation. We exploit sparse decomposition of a single source in order to extract it from multidimensional sensor data, in applications where a rough template of the source is known. This leads to a convex optimization problem, which is solved by a Newton-type method. Complete and overcomplete dictionaries are considered. Simulations with synthetic evoked responses mixed into natural 122-channel MEG data show significant improvement in accuracy of signal restoration.

1. INTRODUCTION

We consider the following problem

$$\mathbf{x}(t) = \mathbf{a}s(t) + \xi(t) \tag{1}$$

where $\mathbf{x}(t)$ is an observed n-channel sensor signal, $s(t)$ is an unknown scalar signal of interest, $\mathbf{a}$ is an unknown n-dimensional vector of weights, and $\xi(t)$ is an n-channel background signal.

We assume that a rough template $\hat{s}(t)$ of the signal $s(t)$ is known in advance. It can be for example, a rectangular pulse, which corresponds to the sign of the original signal, or of its most significant part.

We also assume that $s(t)$ has a sparse representation by means of its decomposition coefficients c_k, obtained in accordance with the signal dictionary of functions $\varphi_k(t)$:

$$s(t) = \sum_{k=1}^{K} c_k\, \varphi_k(t). \tag{2}$$

The functions $\varphi_k(t)$ are called *atoms* or *elements* of the dictionary. These elements do not have to be linearly independent, and instead may form an overcomplete dictionary. Important examples are wavelet and wavelet-related dictionaries (wavelet packets, stationary wavelets, Gabor-type frames, *etc.*, see for example [1, 2, 3] and references therein), or learned dictionaries [4, 5].

Sparsity means that only a small number of coefficients c_k differ significantly from zero. It was shown in [6, 7, 8, 9] that use of sparseness often yields much better blind source separation than other techniques. In this work we use the same property of sparseness for extraction of a single source.

There are other approaches of a single source extraction. For example, Fast ICA algorithm [10] permits the extraction sources from mixtures sequentially, using an approximation of entropy as a criterion for separation. It will not necessary extract the source of interest first, especially when the number of data channels is large. In order to deal with this problem, it was suggested in [11] to initialize separation weights in Fast ICA using a second order method based on maximal correlation with a template. This approach improves the order of source extraction, but it does not exploit the knowledge of a template at the second stage of separation with Fast ICA.

In our work we combine the prior knowledge about the sparsity of a source representation with the knowledge of its template into one optimization objective. Resulting optimization problem is convex (unlike problems arising in usual ICA). It leads to high-quality solution even when the number of data channels is high and total number of samples is small. In our simulations we use 512 samples of 122 channel MEG data. In this situation standard ICA techniques can not give a meaningful separation, because the number of free parameters in the separation 122x122 matrix is much larger than the number of data samples (normally the amount of data used for blind separation of such a data by standard methods is of order 10^5 samples or more, see for example [12]).

In the sequel we will use a matrix notation. Let $t = 1, 2, \ldots, T$ be a discrete time under consideration, $\mathbf{X}$ be a matrix $T \times n$, with discrete signals $x_i(t)$ in its

Supported in part by the Ollendorff Minerva Center and by the Israeli Ministry of Science

columns, and Φ be a matrix $T \times K$ with columns $\varphi_k(t)$. Then, instead of (2), we have

$$\mathbf{s} = \Phi \mathbf{c} \ . \tag{3}$$

If an estimate $\tilde{s}(t)$ of the signal would be known, it could be sparsely decomposed in the dictionary Φ using the following optimization [1]

$$\min_c \|\tilde{\mathbf{s}} - \Phi \mathbf{c}\|^2 + \mu \sum_{k=1}^{K} h(c_k). \tag{4}$$

Here $h(c)$ can be considered as a penalty for non-sparseness. A reasonable choice of $h(c)$ [13, 5] is

$$h(c) = |c|^{1/\gamma}; \qquad \gamma \geq 1, \tag{5}$$

or a smooth approximation thereof. Here we will use a family of convex smooth approximations to the absolute value [6]

$$\begin{aligned} h_1(c) &= |c| - \log(1 + |c|) && (6) \\ h_\alpha(c) &= \alpha h_1(c/\alpha), && (7) \end{aligned}$$

with α being a proximity parameter: $h_\alpha(c) \to |c|$ as $\alpha \to 0^+$. Other approximations to the absolute value can be used as well. For example

$$h_\alpha(c) = \sqrt{c^2 + \alpha}$$

2. SECOND ORDER SOURCE EXTRACTION USING CORRELATION WITH A TEMPLATE

In this section we present a standard approach of maximum correlation with a template, which will be used as a reference point. We look for an estimate of the signal $s(t)$ as a linear combination of the sensor signals

$$\tilde{s}(t) = \sum_i w_i x_i(t) \ , \tag{8}$$

which in a matrix form is

$$\tilde{\mathbf{s}} = \mathbf{X}\mathbf{w} \ , \tag{9}$$

where $\mathbf{w}$ is a vector of weights that we would like to determine.

Suppose that we have an approximate template $\hat{\mathbf{s}}$ of the signal $\mathbf{s}$. Then one can find an estimate of the signal $\mathbf{s}$ in the form (9), which has maximal correlation with the template

$$\max_{\tilde{\mathbf{s}}} \frac{\hat{\mathbf{s}}^T \tilde{\mathbf{s}}}{\|\hat{\mathbf{s}}\| \cdot \|\tilde{\mathbf{s}}\|} \ .$$

It can be rewritten equivalently as

$$\begin{aligned} &\min_{\tilde{\mathbf{s}}} && \|\tilde{\mathbf{s}}\|^2 \\ &\text{Subject to} && \hat{\mathbf{s}}^T \tilde{\mathbf{s}} = 1. \end{aligned} \tag{10}$$

Combining this with (9), we obtain

$$\begin{aligned} &\min_{\mathbf{w}} && \|\mathbf{X}\mathbf{w}\|^2 \\ &\text{Subject to} && \hat{\mathbf{s}}^T \mathbf{X}\mathbf{w} = 1. \end{aligned} \tag{11}$$

The problem can be solved using the method of Lagrange multipliers, which yields

$$\tilde{\mathbf{w}} = \lambda \mathbf{R}_{xx}^{-1} \mathbf{X}^T \hat{\mathbf{s}}, \tag{12}$$

where $\mathbf{R}_{xx}$ is the covariance matrix: $\mathbf{R}_{xx} = \mathbf{X}^T \mathbf{X}$.

3. SPARSE ESTIMATION WITH A TEMPLATE

Suppose now that we have the following two priors:

- sparsity of the coefficients in the representation (3);
- an approximate template $\hat{\mathbf{s}}$ of the signal.

We look for a signal $\tilde{\mathbf{s}}$ with the sparsest representation $\mathbf{c}$ according to the dictionary Φ, which has a unit covariance with the template. In the general case of overcomplete dictionary this leads to the following optimization problem

$$\begin{aligned} &\min_c && \|\tilde{\mathbf{s}} - \Phi \mathbf{c}\|^2 + \mu \sum_{k=1}^{K} h(c_k), \\ &\text{Subject to} && \hat{\mathbf{s}}^T \tilde{\mathbf{s}} = 1. \end{aligned} \tag{13}$$

In the framework of linear estimation (9) we obtain

$$\begin{aligned} &\min_{\mathbf{c},\mathbf{w}} && \|\mathbf{X}\mathbf{w} - \Phi \mathbf{c}\|^2 + \mu \sum_{k=1}^{K} h(c_k), \\ &\text{Subject to} && \hat{\mathbf{s}}^T \mathbf{X}\mathbf{w} = 1. \end{aligned} \tag{14}$$

When the dictionary is *complete*, we obtain significant simplification of the problem: the matrix Φ is invertible, and the coefficients can be estimated directly

$$\tilde{\mathbf{c}} = \Phi^{-1} \tilde{\mathbf{s}} = \Phi^{-1} \mathbf{X}\mathbf{w}. \tag{15}$$

Combining this with (14), where the first term $\|\mathbf{X}\mathbf{w} - \Phi \mathbf{c}\|^2$ vanishes, and using the transformed sensor data

$$\mathbf{Y} = \Phi^{-1} \mathbf{X},$$

we get

$$\begin{array}{ll} \min\limits_{\mathbf{w}} & \sum\limits_{k=1}^{K} h((\mathbf{Y}\mathbf{w})_k), \\ \text{Subject to} & \hat{\mathbf{s}}^T\mathbf{X}\mathbf{w} = 1 \ . \end{array} \qquad (16)$$

Using the method of Lagrange multipliers, we come to

$$\min_{\mathbf{w}} \sum_{k=1}^{K} h((\mathbf{Y}\mathbf{w})_k) - \lambda \hat{\mathbf{s}}^T\mathbf{X}\mathbf{w} \ . \qquad (17)$$

There is a potential for instability in (17): growth of the first term in any direction is asymptotically linear, therefore the minimum of the objective function can be $-\infty$, when λ is too large, and the second term decreases faster than the first term grows. In order to avoid this, we use a monotonic convex transformation of the second term $u(\hat{\mathbf{s}}^T\mathbf{X}\mathbf{w})$, where $u(\cdot)$ is a convex monotonically decreasing function of one variable. For example, we can use quadratic-logarithmic function [14]

$$u_\tau(t) = \begin{cases} \frac{1}{2}t^2 - t, & t \le \tau \\ -(1-\tau)^2 \log\left(\frac{1-2\tau+t}{1-\tau}\right) - \tau + \frac{1}{2}\tau^2 \quad , & t > \tau \ . \end{cases}$$

where $0 \le \tau < 1$. The second derivative of this function *is continuous and bounded* $\forall t \in I\!\!R$. Thus, we are in good position for the Newton minimization. Finally our function for optimization becomes:

$$F(\mathbf{w}) = \sum_{k=1}^{K} h((\mathbf{Y}\mathbf{w})_k) + \lambda u(\hat{\mathbf{s}}^T\mathbf{X}\mathbf{w}) \ . \qquad (18)$$

It is easy to see from the optimality conditions, that (18) yields the same solution $\bar{\mathbf{w}}$ as (17), when λ is changed by a factor of $u'(\hat{\mathbf{s}}^T\mathbf{X}\bar{\mathbf{w}})$.

4. COMPUTATIONAL EXPERIMENTS WITH SYNTHETIC EVOKED RESPONSES MIXED INTO NATURAL MEG RECORDINGS

In order to verify the method, we synthesized a typical evoked brain response and mixed it linearly (with random weights) into real 122-channel MEG recording taken at the rate of 256 *samples/second*. The evoked response (Fig. 1, top plot) is composed of a narrow positive Gaussian pulse with a standard deviation of 4 samples and a wide negative Gaussian pulse with a standard deviation of 10 samples. The second pulse is delayed by 20 samples with respect to the first and decreased in amplitude by a factor 0.6. Other plots in Fig. 1 show few MEG channels already mixed with our synthetic evoked response. As one can see, the response is almost invisible on the background of brain activity. As a template (Fig. 2) we used a rectangular signal, corresponding to the time interval, when the response is above of 10% of its maximal positive value.

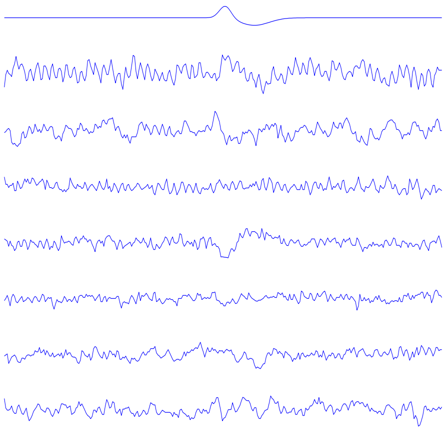


Figure 1: Plot at the top shows synthetic evoked response; other plots show some of MEG channels already mixed with the evoked response: the response is almost invisible on the background of brain activity.

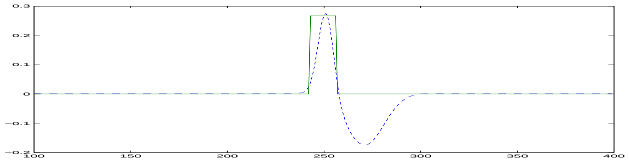
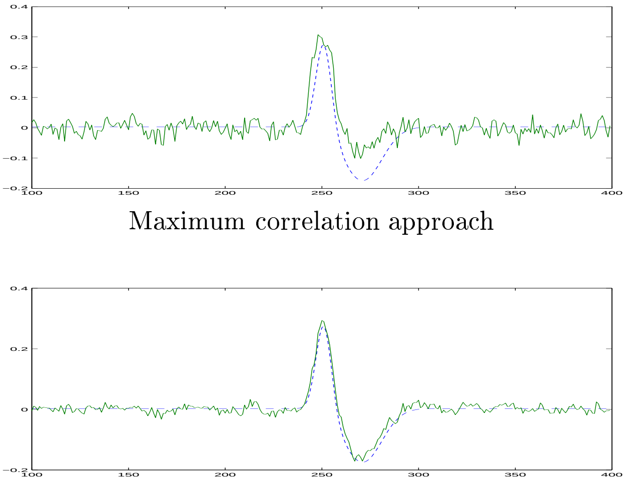


Figure 2: A template (solid line) corresponding to the time interval, when the response (dashed line) is above of 10% of its maximal value.

We compared two methods of recovering evoked responses: the maximum correlation method (12) and our sparse estimation method, which consists in minimization of the objective function (18). We used a wavelet basis Φ with the mother-wavelet *Symmlet-8*, which has eight vanishing moments. This basis is convenient for approximation of smooth functions, like evoked responses are (see for example [2]).

In (18) we used the parameter $\lambda = 1000$, and in (7) the parameter $\alpha = 0.01$. As we observed empirically, a change of the parameters by a factor 10 up and down, does not affect results significantly. Slight improvement of quality can be observed when λ grows and α decreases more significantly, but the problem becomes more difficult for optimization.

As a minimization procedure we used the Newton method with frozen Hessian (see for example [15]). At each iteration the Hessian matrix was computed and



Maximum correlation approach

Sparse estimation

Figure 3: Top plot, solid line - evoked response recovered by using the maximum correlation approach. Bottom plot, solid line - evoked response recovered by using sparse estimation. Dashed line in both plots - the original signal.

three consequent Newton steps were then produced by substituting current gradients into the same Cholessky decomposition of the Hessian (expressions of gradient and Hessian are presented in the Appendix). Cubic linesearch with bisection safeguard and early stopping by Goldstain criterion was used at every Newton step.

The results of reconstruction by the maximum correlation method are shown on the top of Fig.3. The signal-to-noise ratio is significantly better than for original sensor data shown in Fig.1, but the form of the pulse is corrupted, especially the negative part, which was not included in the template.

In Fig.3, bottom, we see the recovered evoked response using our sparse estimation (18). It resembles the original pulse much more accurately, than the maximum correlation method does. Results of 50 simulated trials with random pulse position and random mixing weights are shown in Table 1. The mean-squared error is about 9 times smaller with our method than with the maximum correlation approach.

We also tested some standard ICA techniques with the same data, but the results were meaningless. This can be easily understood taking into account a very small amount of data compared with the number of channels.

5. CONCLUSIONS

Method	Mean-squared error	Std. deviation of sq. error
Max correlation	0.38	0.0354
Sparse estimation	0.044	0.0185

Table 1: Results of 50 simulated trials with random pulse position and random mixing weights

The proposed new approach to extraction a source from multichannel data, using a template and sparse representability of the source in a signal dictionary is most suitable for physiological and medical, as well as wide range of other applications. Our simulations with complete dictionary demonstrate significant superiority of the method over the maximum correlation approach.

A more extensive study has yet to be conducted using overcomplete representations (14), which are more sparse, but also more expensive computationally.

The optimization problems (14) and (16) can be also reformulated as a quadratic or linear programming problems, when $h(\cdot)$ is exactly the absolute value function. This can be done in the spirit of the previous studies [1, 6]. It provides a possibility of using the polynomial complexity algorithms, like *Interior Point Methods*. One can use also a special *Augmented Lagrangian* method for *sum-max* optimization problems [16], which reduces twice the number of variables as compared to the quadratic/linear programming approach, and provides better accuracy of solution. Practical comparison of all these approaches remains open for future research.

6. APPENDIX. GRADIENT AND HESSIAN OF THE OBJECTIVE FUNCTION

Here we obtain derivatives of the objective function (18). Denoting $\mathbf{s} = \mathbf{Y}\mathbf{w}$, $\mathbf{z} = \mathbf{X}^{\mathbf{T}}\hat{\mathbf{s}}$, and $r = \mathbf{z}^T\mathbf{w}$, expression (18) becomes

$$F(\mathbf{w}) = \sum_{k=1}^{K} h(s_k) + \lambda u(r) \qquad (19)$$

Derivation of the gradient formula

Let $\mathbf{h}'(\mathbf{s})$ denotes the vector-column of the first derivatives $h'(s_k)$, the differential of the objective is

$$dF(\mathbf{w}) = \mathbf{ds}^T\mathbf{h}'(\mathbf{s}) + \lambda u'(r)dr\ . \qquad (20)$$

Recalling that $\mathbf{ds} = \mathbf{Y}\mathbf{dw}$ and $dr = \mathbf{z}^T\mathbf{dw} = \mathbf{dw}^T\mathbf{z}$, we get

$$dF(\mathbf{w}) = \mathbf{dw}^T\Big(\mathbf{Y}^T\mathbf{h}'(\mathbf{s}) + \lambda u'(r)\mathbf{z}\Big)\ . \qquad (21)$$

Let $\mathbf{g}$ denotes the gradient of F. Comparing (21) with

$$dF(\mathbf{w}) = \mathbf{g}^T\mathbf{dw} = \mathbf{dw}^T\mathbf{g} \ ,$$

we obtain finally

$$\mathbf{g}(\mathbf{w}) = \mathbf{Y}^T\mathbf{h}'(\mathbf{s}) + \lambda u'(r)\mathbf{z} \ . \qquad (22)$$

Derivation of the Hessian formula

Let Diag $\mathbf{h}''(\mathbf{s})$ denotes the diagonal matrix of the second derivatives $h''(s_k)$. It is easy to obtain from (22)

$$\mathbf{dg}(\mathbf{w}) = \mathbf{Y}^T\text{Diag } \mathbf{h}''(\mathbf{s})\mathbf{ds} + \lambda u''(r)\mathbf{z}dr \ . \qquad (23)$$

Taking into account that $\mathbf{ds} = \mathbf{Ydw}$ and $dr = \mathbf{z}^T\mathbf{dw}$, we get

$$\mathbf{dg}(\mathbf{w}) = \mathbf{Y}^T\text{Diag } \mathbf{h}''(\mathbf{s})\mathbf{Ydw} + \lambda u''(r)\mathbf{zz}^T\mathbf{dw} \ . \quad (24)$$

Comparing this with the known expression

$$\mathbf{dg}(\mathbf{w}) = \mathbf{H}(\mathbf{w})\mathbf{dw},$$

where $\mathbf{H}(\mathbf{w})$ is a Hessian matrix, we finally obtain

$$\mathbf{H}(\mathbf{w}) = \mathbf{Y}^T\text{Diag } \mathbf{h}''(\mathbf{s})\mathbf{Y} + \lambda u''(r)\mathbf{zz}^T \ .$$

7. REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," 1996. http://www-stat.stanford.edu/˜donoho/Reports/.

[2] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[3] M. Zibulski and Y. Y. Zeevi, "Analysis of multiwindow gabor-type schemes by frame methods," *Applied and Computational Harmonic Analysis*, vol. 4, pp. 188–221, 1997.

[4] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[5] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[6] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.

[7] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using the sparsity of the short-time fourier transform," in *International Workshop on Independent Component Analysis and Blind Signal Separation*, (Helsinki, Finland), June 19–20 2000.

[8] P. Kisilev, M. Zibulevsky, Y. Y. Zeevi, and B. A. Pearlmutter, "Multiresolution framework for sparse blind source separation," tech. rep., Department of Electrical Engineering, Technion. Haifa, Israel, 2000. http://ie.technion.ac.il/˜mcib/.

[9] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition in a signal dictionary," in *Independent Components Analysis: Princeiples and Practice* (S. J. Roberts and R. M. Everson, eds.), Cambridge University Press, 2001.

[10] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[11] A. K. Barros, R. Vigario, V. Jousmaki, and N. Ohnishi, "Extraction of event-related signals from multi-channel bioelectrical measurements," *IEEE Trans. Biomed. Eng.*, 2001. To appear.

[12] A. C. Tang, B. A. Pearlmutter, and M. Zibulevsky, "Blind separation of multichannel neuromagnetic responses," in *Computational Neuroscience*, pp. 1115–1120, 1999. Published in a special issue of *Neurocomputing* volume 32–33 (2000).

[13] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[14] A. Ben-Tal and M. Zibulevsky, "Penalty/barrier multiplier methods for convex programming problems," *SIAM Journal on Optimization*, vol. 7, no. 2, pp. 347–366, 1997.

[15] L. Mosheyev and M. Zibulevsky, "Penalty/barrier multiplier algorithm for semidefinite programming," *Optimization Methods and Software*, vol. 13, no. 4, pp. 235–261, 2000.

[16] M. Zibulevsky, *Penalty/Barrier Multiplier Methods for Large-Scale Nonlinear and Semidefinite Programming.* PhD thesis, Technion – Israel Institute of Technology, 1996. http://ie.technion.ac.il/˜mcib/.