Learning Recurring Patterns in Large Signals with Convolutional Dictionary Learning

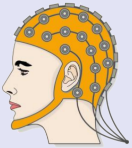
Thomas Moreau
Parietal – INRIA Saclay

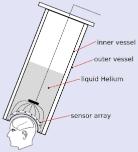
Joint work with T. Dupré la Tour (Telecom), M. Jas (Telecom), A. Gramfort (INRIA), N. Vayatis (CMLA), L. Oudre (UP13)

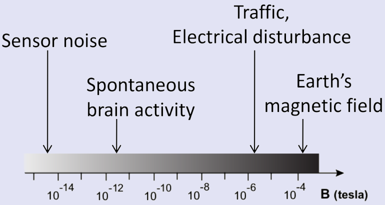
inventors for the digital world

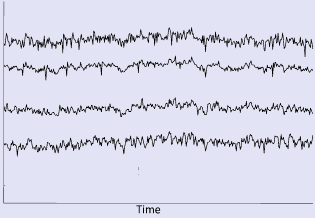
Studying brain activity through electromagnetic signals

- Brain (electrical) activity produces an electromagnetic field.
- This can be measured with EEG or MEG.









Goal: Study Oscillation in Neural Data

Oscillations are believed to play an important role in cognitive functions.

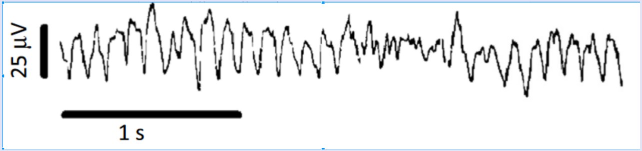
Many studies rely on Fourier or wavelet analyses:

- Easy interpretation,
- Standard analysis *e.g.* canonical bands alpha, beta or theta.

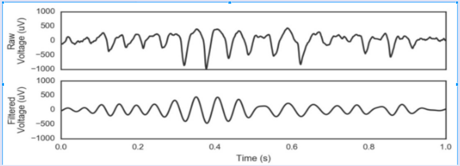
[Buzsaki, 2006]

Goal: Study Oscillation in Neural Data

However, some brain rhythms are not sinusoidal, *e.g.*mu-waves [Hari, 2006]

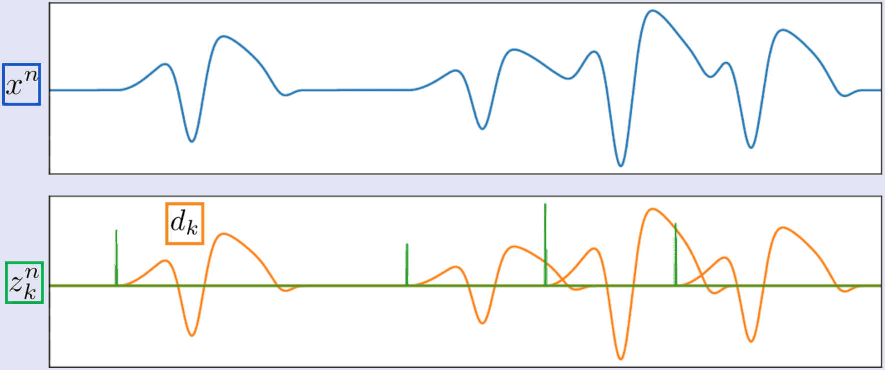


and filtering degrades waveforms



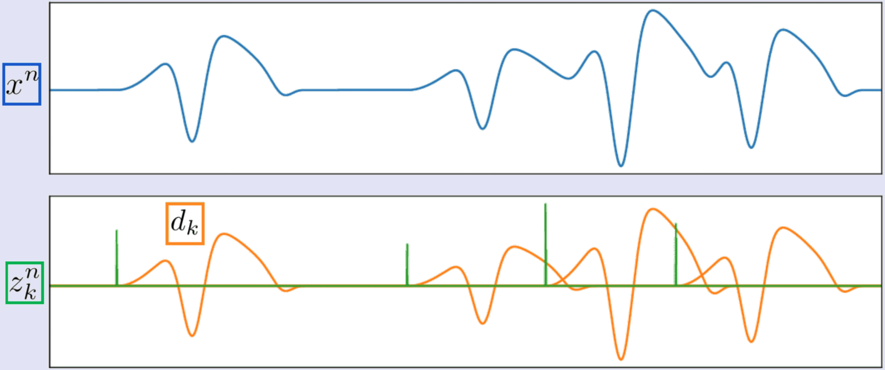
$\Rightarrow$ Can we do better with data-driven approach?

Key idea: decouple the localization of the patterns and their shape



Extracting shift invariant patterns

Key idea: decouple the localization of the patterns and their shape

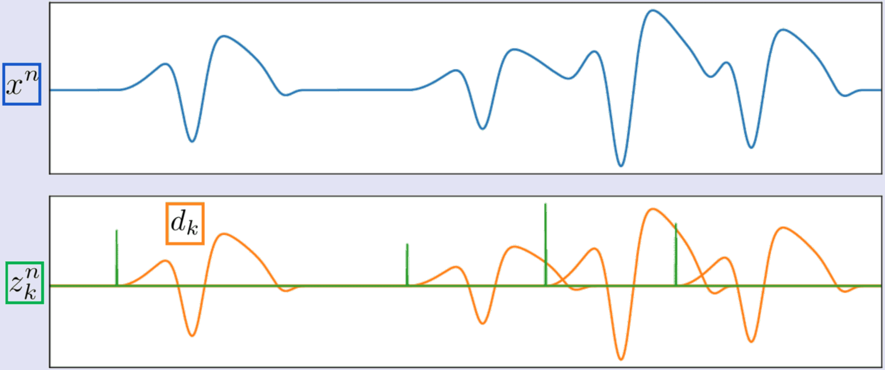


Convolutional Representation:

$$\boxed{x^n}[t] = \sum_{k=1}^{K} (\boxed{z_k^n} * \boxed{d_k})[t] + \varepsilon[t]$$

Extracting shift invariant patterns

Key idea: decouple the localization of the patterns and their shape



Convolutional Dictionary Learning:

$$\min_{d,z} \sum_{n=1}^{N} \frac{1}{2} \left\| \boxed{x^n} - \sum_{k=1}^{K} \boxed{z_k^n} * \boxed{d_k} \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\boxed{z_k^n}\|_1,$$

$$\text{s.t.} \quad \|\boxed{d_k}\|_2^2 \leq 1$$

Shift-invariant Patterns in images

Images also have shift-invariant patterns that we might want to detect.

Convolutional Dictionary Learning (CDL) [Grosse et al., 2007]

For a set of N univariate signals x^n, solve

$$\min_{d_k, z_k^n} \sum_{n=1}^{N} \frac{1}{2} \|x^n - \sum_{k=1}^{K} z_k^n * d_k\|_2^2 + \lambda \sum_{k=1}^{K} \|z_k^n\|_1 \tag{1}$$

Hypothesis: patterns d_k are not present everywhere in the signal. They are localized in time.

$\Rightarrow$ Sparse activation signals z

Extra hypothesis: the patterns are in the ℓ_2-ball: $\|d_k\|_2^2 \leq 1$.

The problem 1 is not jointly convex in z_k^n, and d_k it is convex in each block of coordinate.

Alternate minimization (*a.k.a.* Bloc Coordinate Descent):

- **Z-step:** given a fixed estimate of the atom, compute the activation signal z_k^n associated to each signal X^n.
- **D-step:** given a fixed estimate of the activation, update the atoms in the dictionary d_k.

Convolutional Sparse Coding with Locally Greedy Coordinate Descent (LGCD)

References

- Moreau, T., Oudre, L., and Vayatis, N. (2018). DICOD: Distributed Convolutional Sparse Coding. In *International Conference on Machine Learning (ICML)*, pages 3626–3634, Stockholm, Sweden. PMLR (80)

Convolutional Sparse Coding

N independent problem such that

$$\min_{z_k^n} E(z^n) = \frac{1}{2}\left\|x^n - \sum_{k=1}^{K} z_k^n * d_k\right\|_2^2 + \lambda \sum_{k=1}^{K} \|z_k^n\|_1 \quad .$$

This problem is convex in z_k and can be solved with different techniques:

- Greedy CD [Kavukcuoglu et al., 2010]
- Fista [Chalasani et al., 2013]
- ADMM [Bristow et al., 2013]
- L-BFGS [Jas et al., 2017]

⇒ These methods can be slow for long signals as the complexity of each iteration is at least linear in the length of the signal.

Greedy coordinate descent (GCD) [Kavukcuoglu et al., 2010]

For the Greedy Coordinate Descent, only 1 coordinate is updated at each iteration:

1. The coordinate $z_{k_0}[t_0]$ is updated to its optimal value $z'_{k_0}[t_0]$ when all other coordinate are fixed:

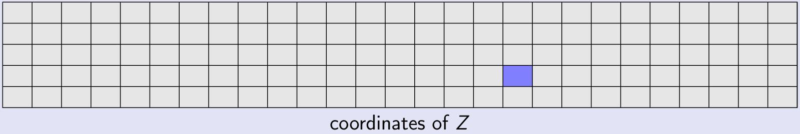
$$z'_k[t] = \max\left(\frac{\beta_k[t] - \lambda}{\|d_k\|_2^2}, 0\right),$$

with $\beta_k[t] = \left[\left(X - \sum_{l=1}^{K} z_l * d_l + z_k[t]e_t * d_k\right) * d_k^{\Lsh}\right][t]$

2. Greedy coordinate selection:

$$(k, t) = \underset{(k,t)}{\operatorname{argmax}} \left| z_k[t] - z'_k[t] \right|$$

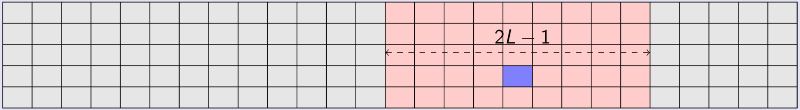
We introduced the LGCD method which is an extension of GCD.



GCD has $\mathcal{O}(KT)$ computational complexity.

Locally greedy coordinate descent (LGCD) [Moreau et al., 2018]

We introduced the LGCD method which is an extension of GCD.

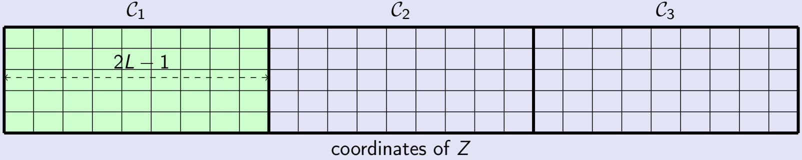


coordinates of Z

GCD has $\mathcal{O}(KT)$ computational complexity.

But the update itself has complexity $\mathcal{O}(KL)$

We introduced the LGCD method which is an extension of GCD.

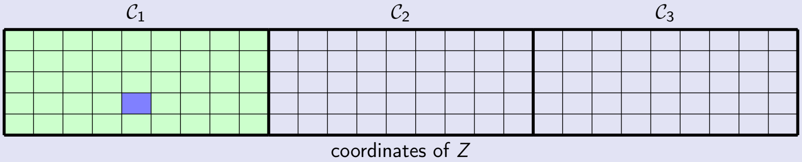


With a partition $\mathcal{C}_m$ of the signal domain $[\![1, K]\!] \times [\![0, T]\!]$,

$$\mathcal{C}_m = [\![1, K]\!] \times [\![\frac{(m-1)\widetilde{T}}{M}, \frac{m\widetilde{T}}{M}]\!]$$

Locally greedy coordinate descent (LGCD) [Moreau et al., 2018]

We introduced the LGCD method which is an extension of GCD.



With a partition $\mathcal{C}_m$ of the signal domain $[\![1, K]\!] \times [\![0, T]\!]$,

$$\mathcal{C}_m = [\![1, K]\!] \times [\![\frac{(m-1)\widetilde{T}}{M}, \frac{m\widetilde{T}}{M}]\!]$$

The coordinate to update is chosen greedily on a sub-domain $\mathcal{C}_m$

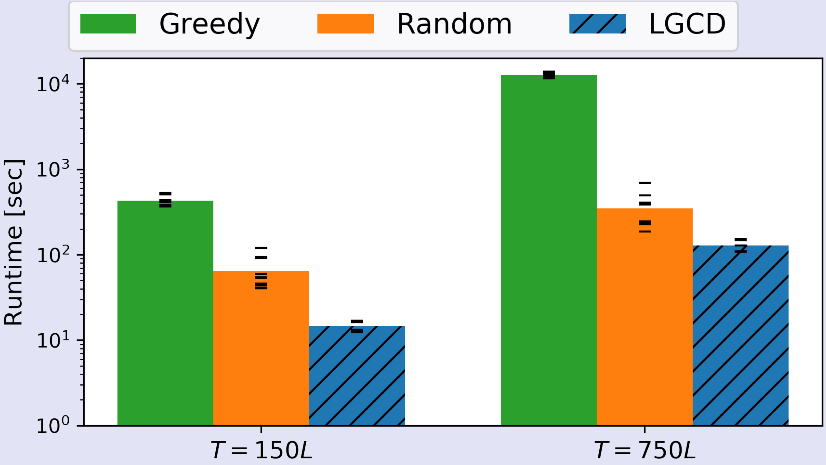
$$\mathcal{O}(\text{Coordinate selection}) = \mathcal{O}(\text{Coordinate Update})$$

The overall iteration complexity is $\mathcal{O}(KL)$ instead of $\mathcal{O}(KT)$.

$$\Rightarrow \text{Efficient for sparse } Z$$

Fast optimization

Comparison of the coordinate selection strategy for CD on simulated signals
We set $K = 10$, $L = 150$, $\lambda = 0.1\lambda_{\text{max}}$



Distributed optimization for CSC

References

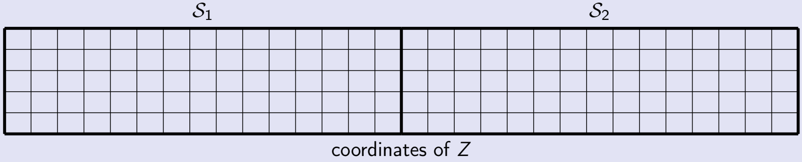
- Moreau, T. and Gramfort, A. (2019). Distributed Convolutional Dictionary Learning (DiCoDiLe): Pattern Discovery in Large Images and Signals. *preprint ArXiv (to be submitted)*

The update of the W coordinates $(k_w, \omega_w)_{w=1}^{W}$ with additive update $\Delta Z_{k_w}[\omega_w]$ changes the cost by:

$$\Delta E = \overbrace{\sum_{i=1}^{W} \Delta E_w}^{\text{iterative steps}} - \underbrace{\sum_{w \neq w'} (d_{k_w} * d^{\Lsh}_{k_{w'}})[\omega_{w'} - \omega_w] \Delta Z_{k_w}[\omega_w] \Delta Z_{k_{w'}}[\omega_{w'}]}_{\text{interference}},$$

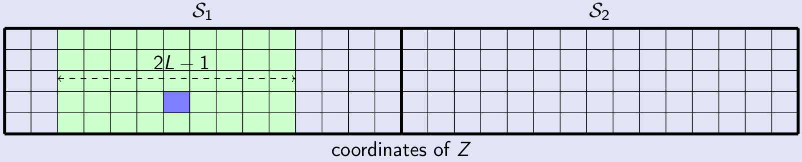
$\Rightarrow$ If the updates are far enough, they can be considered as independent.

Distributed Convolutional Coordinate Descent (DICOD)



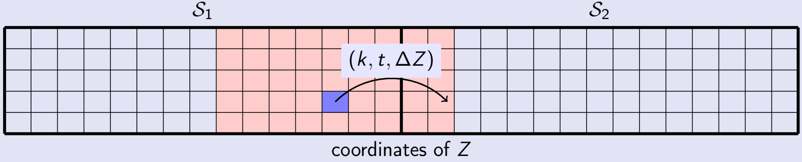
- Split the coordinates in continuous sub-segment $\mathcal{S}_w = \left[\frac{(w-1)T}{W}, \frac{wT}{W}\right[$.

Distributed Convolutional Coordinate Descent (DICOD)



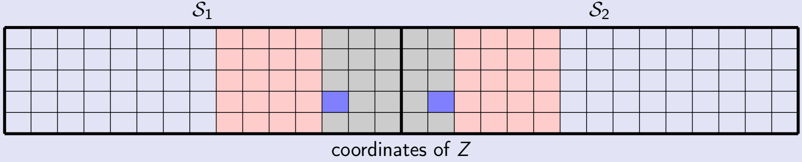
- Split the coordinates in continuous sub-segment $\mathcal{S}_w = \left[\frac{(w-1)T}{W}, \frac{wT}{W}\right[$.
- Use Greedy updates in parallel in each sub-segment.

Distributed Convolutional Coordinate Descent (DICOD)



- Split the coordinates in continuous sub-segment $\mathcal{S}_w = \left[\frac{(w-1)T}{W}, \frac{wT}{W}\right[$.
- Use Greedy updates in parallel in each sub-segment.
- Notify neighbor workers when the update is on the border of $\mathcal{S}_w$.

Distributed Convolutional Coordinate Descent (DICOD)



- Split the coordinates in continuous sub-segment $\mathcal{S}_w = \left[\frac{(w-1)T}{W}, \frac{wT}{W}\right[$.
- Use Greedy updates in parallel in each sub-segment.
- Notify neighbor workers when the update is on the border of $\mathcal{S}_w$.

This algorithm converges to the solution of the CSC for 1D signals but not for higher dimension signals such as images.

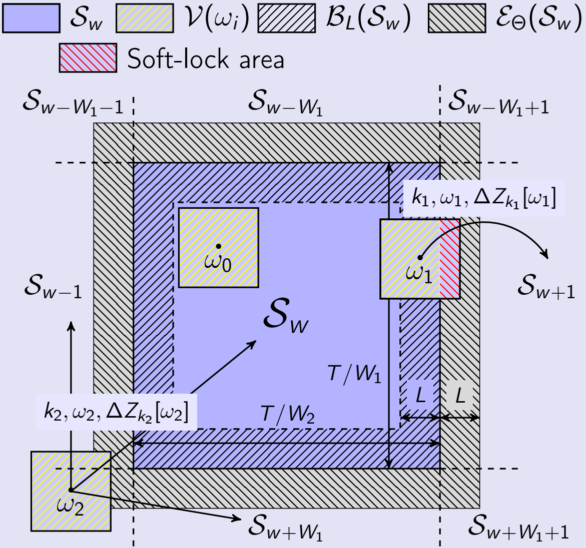
Distributed Convolutional Dictionary Learning (DiCoDiLe-Z)

- Extension of DICOD for high dimensional signals.
- Use LGCD locally in each workers (better iteration complexity).
- Use Soft-locks to avoid interference (ensure convergence).

Distributed Convolutional Dictionary Learning (DiCoDiLe-Z)

- Update candidate ω_0 is independent of other workers as

$$\mathcal{V}(\omega_0) \subset \mathcal{S}_w$$

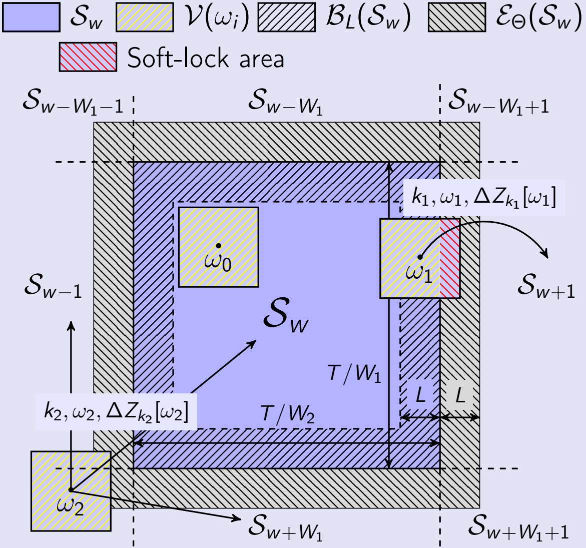


Distributed Convolutional Dictionary Learning (DiCoDiLe-Z)

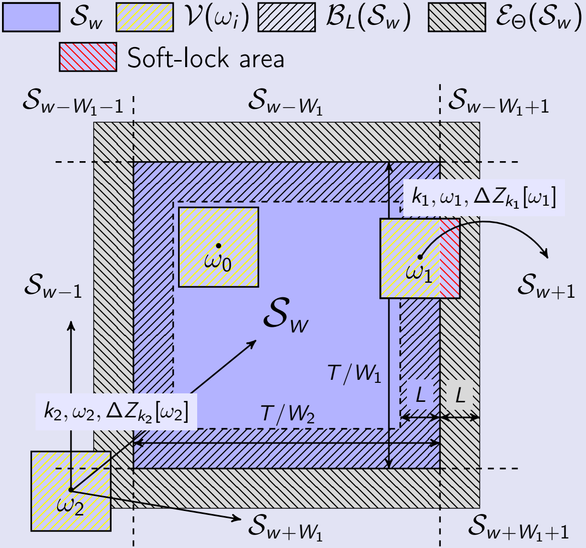
- Update candidate ω_1 impacts $\mathcal{S}_{w+1}$

$$\mathcal{V}(\omega_1) \not\subset \mathcal{S}_w$$

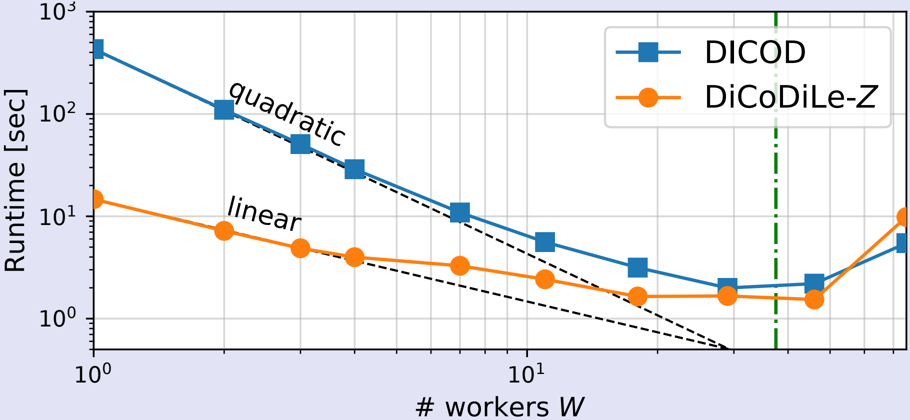
- It is accepted only is no better update is possible in the "soft-locked" area.
- Need to notify $\mathcal{S}_{w+1}$.



- Updates in ω_2 need to notify worker w to maintain consistent estimate in the border zone $\mathcal{B}_L(\mathcal{S}_w)$.
- Low communication: decentralized and below 1ko.



Numerical speed-up



Running time as a function fo the number of workers W.

Rank-1 Constrained Convolutional Dictionary Learning

References

- Dupré la Tour, T., Moreau, T., Jas, M., and Gramfort, A. (2018). Multivariate Convolutional Sparse Coding for Electromagnetic Brain Signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3296–3306, Montreal, Canada

D-step: solving for the atoms

The dictionary update is performed by minimizing

$$\min_{\|d_k\|_2 \leq 1} E(\{d_k\}_k) \triangleq \sum_{n=1}^{N} \frac{1}{2} \|X^n - \sum_{k=1}^{K} z_k^n * d_k\|_2^2 \quad . \tag{2}$$

Computing $\nabla_{d_k} E(\{d_k\}_k)$ can be done efficiently

$$\nabla_{d_k} E(\{d_k\}_k) = \sum_{n=1}^{N} (z_k^n)^{\Lsh} * \left(x^n - \sum_{l=1}^{K} z_l^n * d_l \right) = \Phi_k - \sum_{l=1}^{K} \Psi_{k,l} * d_l \ ,$$

$\Rightarrow$ Save with Projected Gradient Descent (PGD) with an Armijo backtracking line-search for the d-step [Wright and Nocedal, 1999].

How to extend CSC to multivariate signals?

We can just use multivariate convolution,

$$\underbrace{X[t]}_{\in\mathbb{R}^P} = \sum_{k=1}^{K} (z_k * \boldsymbol{D}_k)[t] = \sum_{k=1}^{K}\sum_{\tau=1}^{L} z_k[t-\tau]\underbrace{\boldsymbol{D}_k[\tau]}_{\in\mathbb{R}^P}$$

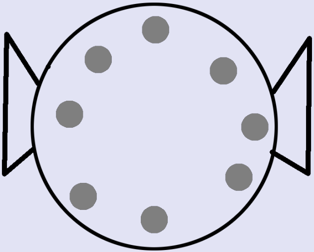
with:

- X a multivariate signal of length T in $\mathbb{R}^P$
- $\boldsymbol{D}_k$ a multivariate signal of length L in $\mathbb{R}^P$
- z_k a univariate activation signal of length $\widetilde{T} = T - L + 1$

However, this model does not account for the physics of the problem.

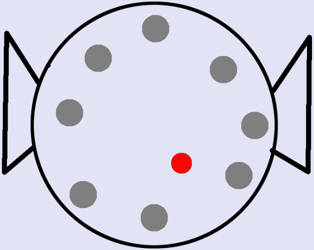
EM wave diffusion

- Recording here with 8 sensors



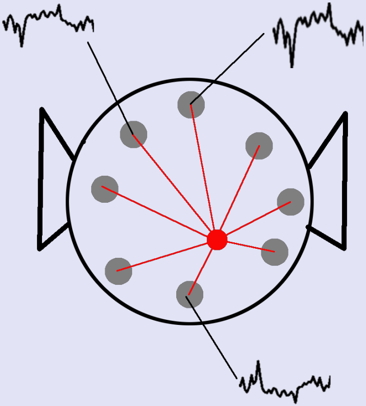
EM wave diffusion

- Recording here with 8 sensors
- EM activity in the brain



EM wave diffusion

- Recording here with 8 sensors
- EM activity in the brain
- The electric field is spread **linearly** and **instantaneously** over all sensors (Maxwell equations)



Multivariate CSC with rank-1 constraint

Idea: Impose a rank-1 constraint on the dictionary atoms D_k

To make the problem tractable, we decided to use auxiliary variables u_k and v_k *s.t.* $D_k = u_k v_k \top$.

$$\min_{u_k, v_k, z_k^n} \sum_{n=1}^{N} \frac{1}{2} \left\| X^n - \sum_{k=1}^{K} z_k^n * (u_k v_k^\top) \right\|_2^2 + \lambda \sum_{k=1}^{K} \|z_k^n\|_1 \, , \quad (3)$$
$$\text{s.t.} \quad \|u_k\|_2^2 \leq 1 \, , \, \|v_k\|_2^2 \leq 1 \text{ and } z_k^n \geq 0 \, .$$

Here,

- $u_k \in \mathbb{R}^P$ is the spatial pattern of our atom
- $v_k \in \mathbb{R}^L$ is the temporal pattern of our atom

Update in u_k and v_k

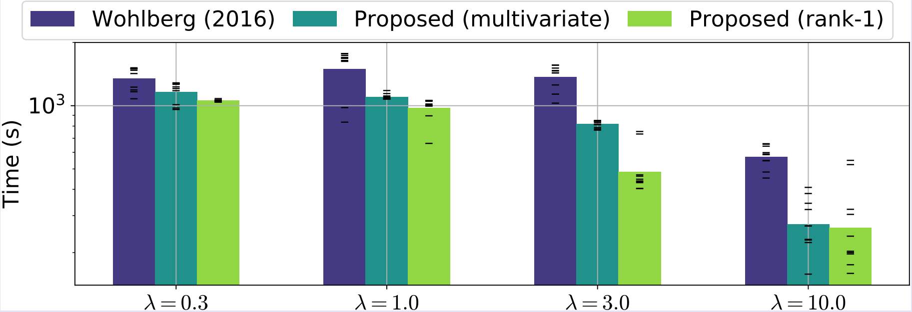
The problem is not jointly convex in u_k and v_k.

Use an alternate minimization on these two blocks.

The gradient can also be computed using sufficient statistics ϕ and ψ:

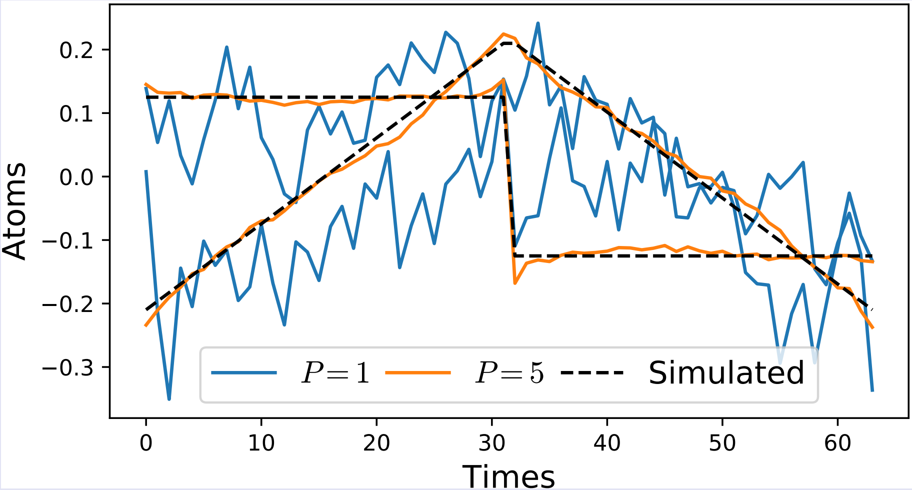
$$\nabla_{u_k} E(\{u_k\}_k, \{v_k\}_k) = \nabla_{D_k} E(\{u_k\}_k, \{v_k\}_k) v_k \quad \in \mathbb{R}^P \ ,$$
$$\nabla_{v_k} E(\{u_k\}_k, \{v_k\}_k) = u_k^\top \nabla_{D_k} E(\{u_k\}_k, \{v_k\}_k) \quad \in \mathbb{R}^L \ ,$$

Comparison with multivariate methods on somato dataset with $T = 134,700$, $K = 8$, $P = 5$ and $L = 128$



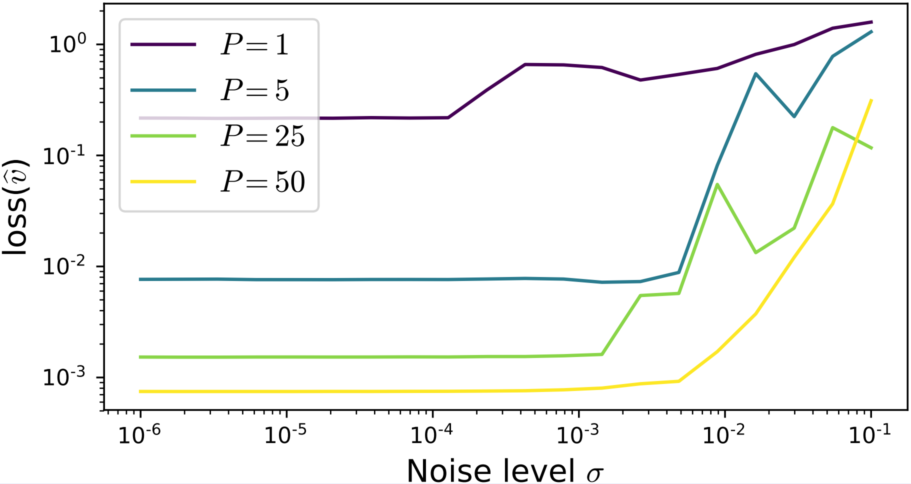
Pattern recovery

Patterns recovered with $P = 1$ and $P = 5$. The signals were generated with the two simulated temporal patterns and with $\sigma = 10^{-3}$.



Pattern recovery

Evolution of the recovery loss with σ for different values of P. Using more channels improves the recovery of the original patterns.

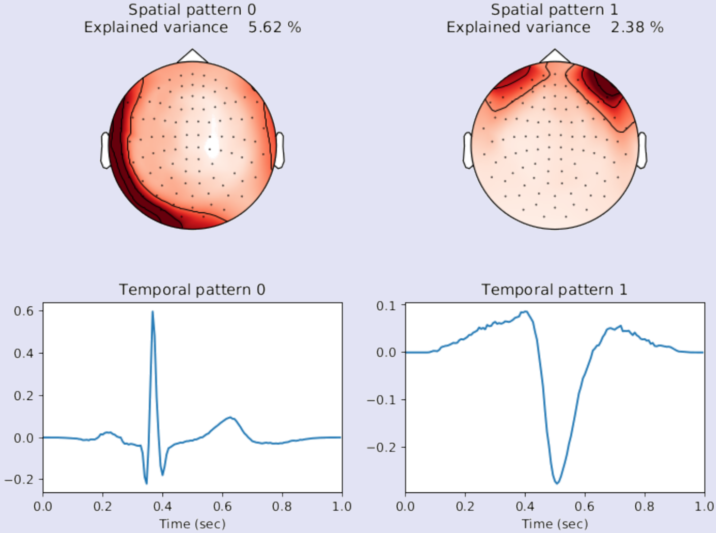


Experiments on REal Data

Good time to wake-up if you got lost in the previous section!

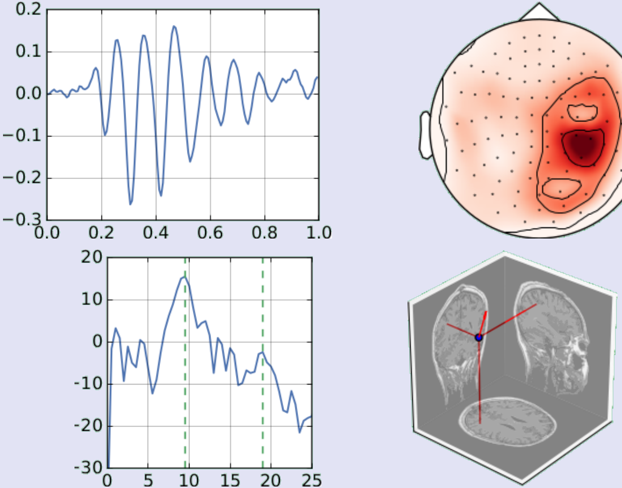
MNE somatosensory data

A selection of temporal waveforms of the atoms learned on the MNE sample dataset.

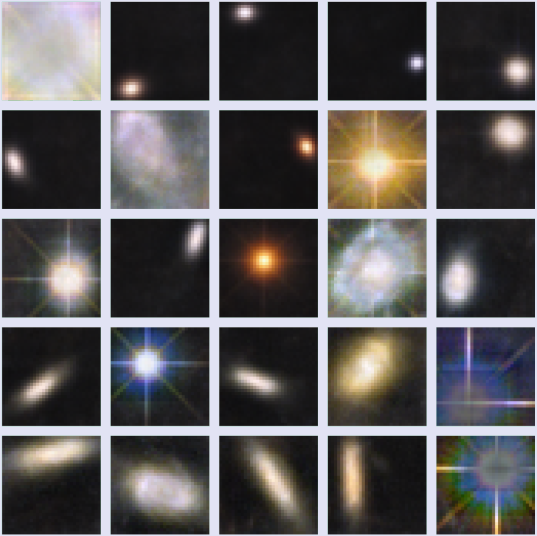


MNE somatosensory data

Atoms revealed using the MNE somatosensory data. Note the non-sinusoidal comb shape of the mu rhythm.



Encoding HST images with CDL



Atoms 32×32 learned with DiCoDiLe on image STScI-H-2016-39-a (resolution 6000×3664).

The atoms are order with $\|Z_k\|_1$.

Conclusion

LGCD and DiCoDile: Efficient algorithm to scale Convolutional Dictionary Learning to large signals.

Rank-1 constraints: Adapt the constraints to the type of patterns researched.

Ahead of us:

- Scale invariant atoms?
- Pattern detection with extra prior:

Thanks!

Code available online:

alphacsc : alphacsc.github.io

DICOD (& DiCoDiLe soon) : github.com/tommoral/dicod

Slides are on my web page:

tommoral.github.io @tomamoral

Reference

Bristow, H., Eriksson, A., and Lucey, S. (2013). Fast convolutional sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 391–398, Portland, OR, USA.

Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

Chalasani, R., Principe, J. C., and Ramakrishnan, N. (2013). A fast proximal method for convolutional sparse coding. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–5, Dallas, TX, USA.

Dupré la Tour, T., Moreau, T., Jas, M., and Gramfort, A. (2018). Multivariate Convolutional Sparse Coding for Electromagnetic Brain Signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3296–3306, Montreal, Canada.

Grosse, R., Raina, R., Kwong, H., and Ng, A. Y. (2007). Shift-Invariant Sparse Coding for Audio Classification. *Cortex*, 8:9.

Hari, R. (2006). Action–perception connection and the cortical mu rhythm. *Progress in brain research*, 159:253–260.

Jas, M., Dupré la Tour, T., Şimşekli, U., and Gramfort, A. (2017). Learning the Morphology of Brain Signals Using Alpha-Stable Convolutional Sparse Coding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–15, Long Beach, CA, USA.

Kavukcuoglu, K., Sermanet, P., Boureau, Y.-l., Gregor, K., and Lecun, Y. (2010). Learning Convolutional Feature Hierarchies for Visual Recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1090–1098, Vancouver, Canada.

Moreau, T. and Gramfort, A. (2019). Distributed Convolutional Dictionary Learning (DiCoDiLe): Pattern Discovery in Large Images and Signals. *preprint ArXiv (to be submitted)*.

Moreau, T., Oudre, L., and Vayatis, N. (2018). DICOD: Distributed Convolutional Sparse Coding. In *International Conference on Machine Learning (ICML)*, pages 3626–3634, Stockohlm, Sweden. PMLR (80).