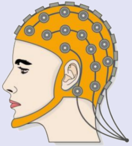
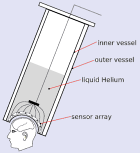
Multivariate Convolutional Sparse Coding for Electromagnetic Brain Signals

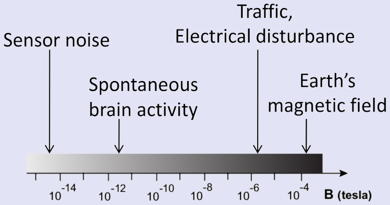
Dupré La Tour T., **TM**, Mainak J., Gramfort A.
INRIA Saclay

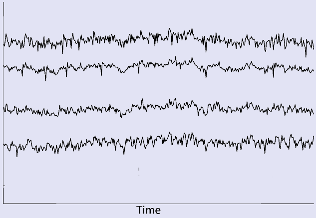
Studying brain activity through electromagnetic signals

- 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.

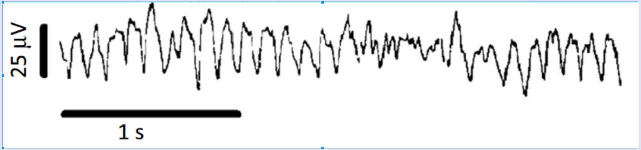
Many studies rely on Fourier or wavelet analyses:

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

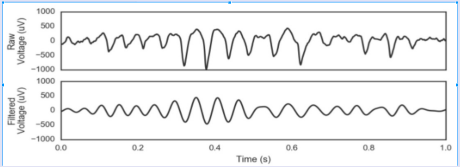
[Buzsáki, 2006]

Goal: Study Oscillation in Neural Data

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



and filtering degrades waveforms

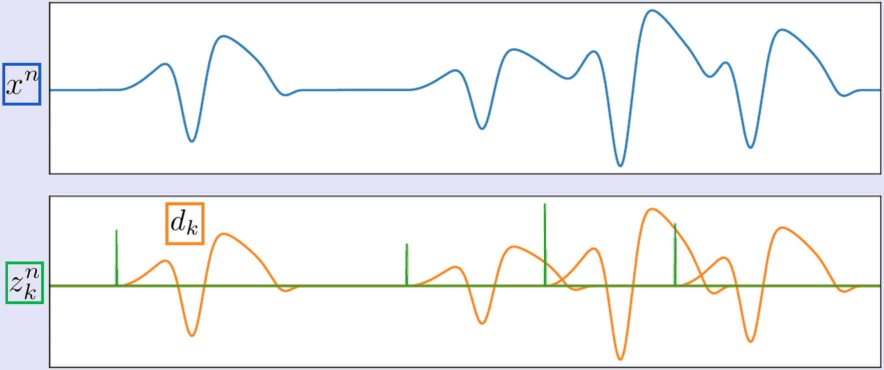


The shape of the waveform can be linked to the information flow between neurons.

$\Rightarrow$ Can extract them with an unsupervised data-driven approach?

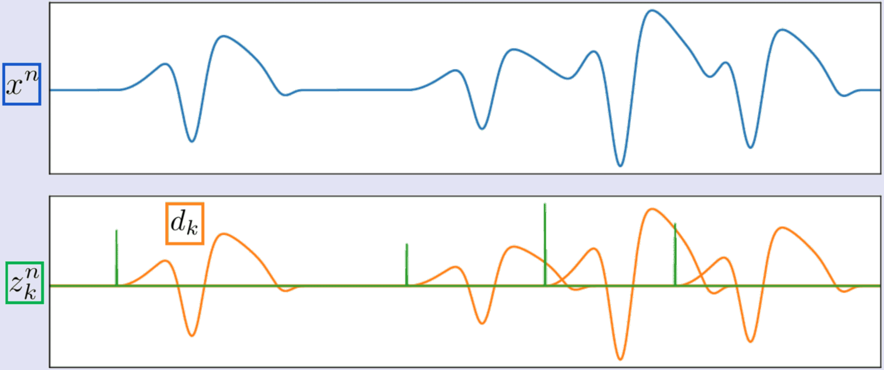
Extracting shift invariant patterns

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

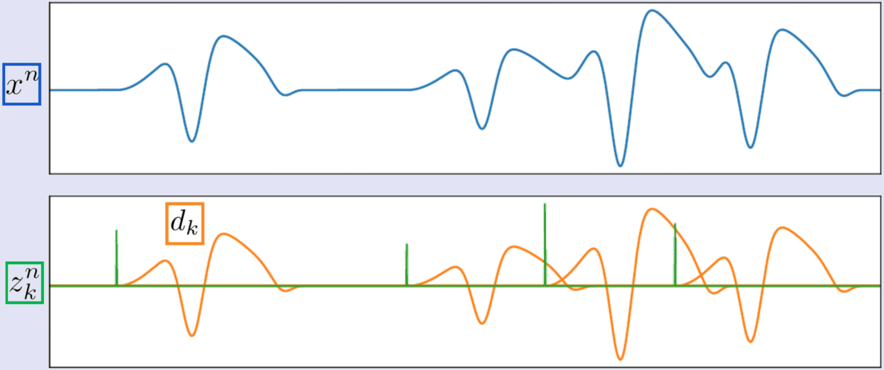


Convolutional Representation:

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

Extracting shift invariant patterns

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



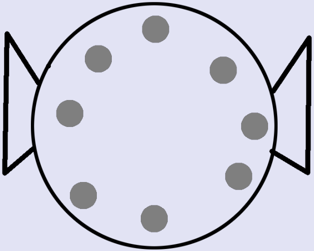
Convolutional Dictionary Learning:

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

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

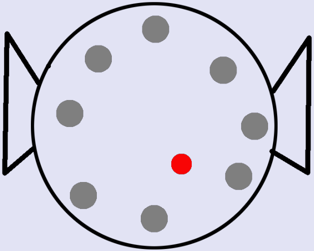
EM wave diffusion

- Recording here with 8 sensors



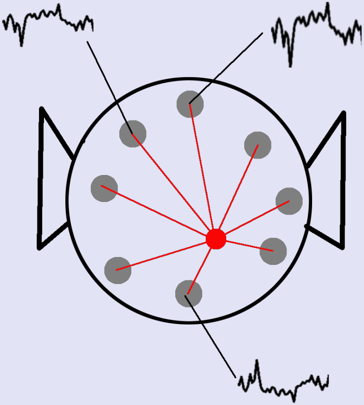
EM wave diffusion

- 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.* $D_k = u_k v_k^\top$.

$$\min_{u_k, v_k, z_k^n} \sum_{n=1}^{N} \frac{1}{2} \left\| X^n - \sum_{k=1}^{K} z_k^n * (u_k v_k^\top) \right\|_2^2 + \lambda \sum_{k=1}^{K} \left\| z_k^n \right\|_1 , \tag{1}$$

$$\text{s.t.} \quad \|u_k\|_2^2 \leq 1 \text{ , } \|v_k\|_2^2 \leq 1 \text{ and } z_k^n \geq 0 \text{ .}$$

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

Optimization strategy

Tri-convex: The problem is not jointly convex in z_k^n, u_k and v_k but it is convex in each block of coordinate.

We can use a block coordinate descent, aka alternate minimization, to converge to a local minima of this problem. The 3 following steps are applied alternatively:

- **Z-step:** given a fixed estimate of the atom, compute the activation signal z_k^n associated to each signal X^n.
- **u-step:** given a fixed estimate of the activation and temporal pattern, update the spatial pattern u_k.
- **v-step:** given a fixed estimate of the activation and spatial pattern, update the temporal pattern v_k.

Z-step: Locally greedy coordinate descent (LGCD)

N independent problem such that

$$\min_{z_k^n \geq 0} \frac{1}{2} \left\| X^n - \sum_{k=1}^{K} z_k^n * D_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \left\| z_k^n \right\|_1 \quad .$$

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.

Z-step: Locally greedy coordinate descent (LGCD)

For the Greedy Coordinate Descent, only 1 coordinate is updated at each iteration: [Kavukcuoglu et al., 2010]

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 $\beta_k[t] = \left[D_k^{\urcorner} * \left(X - \sum_{l=1}^K z_l * D_l + z_k[t]e_t * D_k\right)\right][t]$

For each coordinate update, it is possible to maintain the value of β with $\mathcal{O}(KL)$ operations.

Z-step: Locally greedy coordinate descent (LGCD)

For the Greedy Coordinate Descent, only 1 coordinate is updated at each iteration: [Kavukcuoglu et al., 2010]

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.

2. The updated coordinate is chosen

- Cyclic selection: $\mathcal{O}(1)$ [Friedman et al., 2007]
- Randomized selection: $\mathcal{O}(1)$ [Nesterov, 2010]
- Greedy selection: $\mathcal{O}(K\widetilde{T})$ [Osher and Li, 2009]
  by maximizing $|z_k[t] - z'_k[t]|$
- Locally Greedy selection: $\mathcal{O}(KL)$ [Moreau et al., 2018]
  by maximizing $|z_k[t] - z'_k[t]|$ on a sub-segment.

Z-step: Locally greedy coordinate descent (LGCD)

For the Greedy Coordinate Descent, only 1 coordinate is updated at each iteration: [Kavukcuoglu et al., 2010]

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.

2. The updated coordinate is chosen

- Cyclic selection: $\mathcal{O}(1)$ [Friedman et al., 2007]
- Randomized selection: $\mathcal{O}(1)$ [Nesterov, 2010]
- Greedy selection: $\mathcal{O}(K\widetilde{T})$ [Osher and Li, 2009] by maximizing $|z_k[t] - z'_k[t]|$
- Locally Greedy selection: $\mathcal{O}(KL)$ [Moreau et al., 2018] by maximizing $|z_k[t] - z'_k[t]|$ on a sub-segment.

D-step: solving for the atoms

The dictionary update is performed by minimizing

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

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

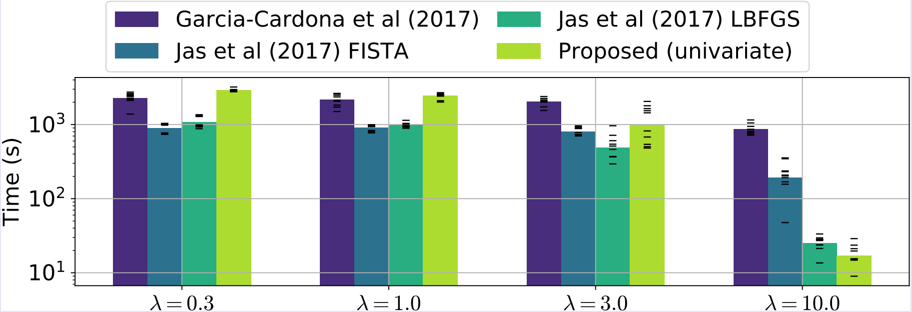
$$\nabla_D E(D) = \sum_{n=1}^{N} (z_k^n)^{\urcorner} * \left(x^n - \sum_{l=1}^{K} z_l^n * D_l \right) = \Phi_k - \sum_{l=1}^{K} \Psi_{k,l} * D_l \ ,$$

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

Experiments

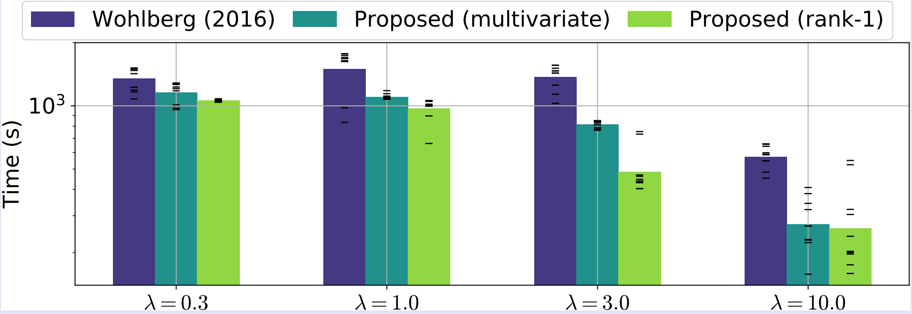
Good time to wake-up if you got lost in the previous section!

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



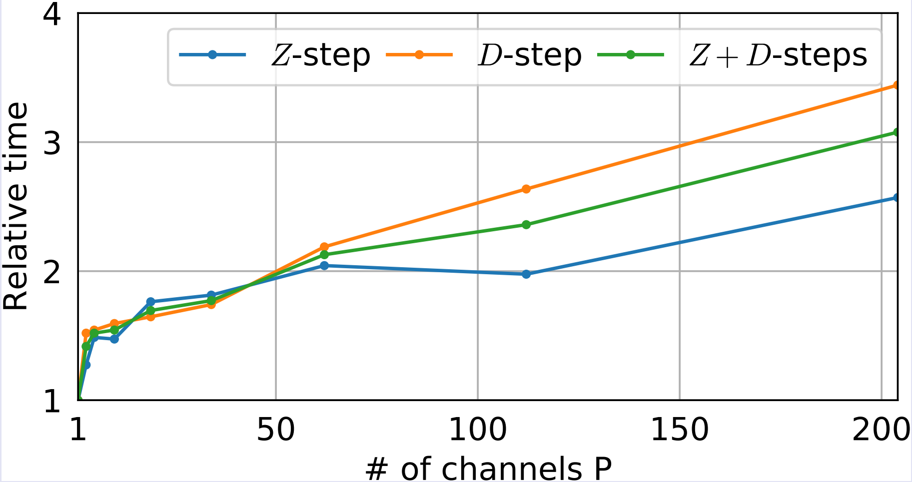
Fast optimization

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



Good scaling in the number of channels P

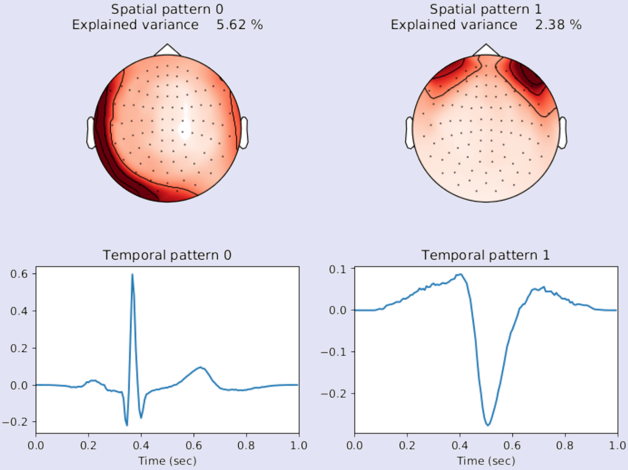
Scaling relative to P on somato dataset with $T = 134,700$, $K = 2$, and $L = 128$



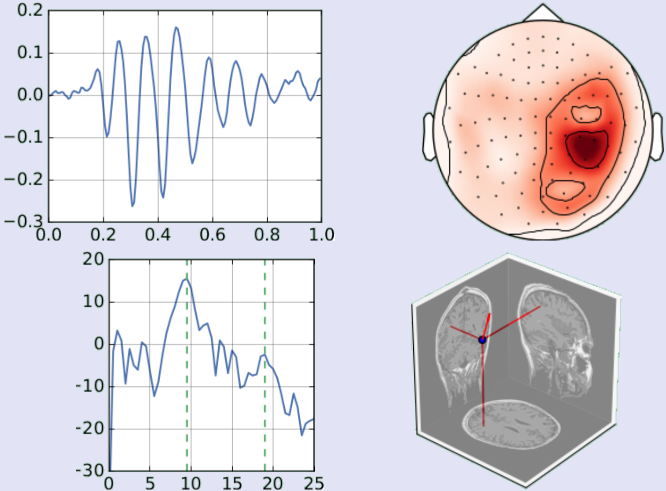
Experiments on MEG data

Even better time to wake-up!

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



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



Conclusion

- We proposed a model for multivariate CSC with rank-1 constraint. This model makes sense for different type of data.
- We proposed a fast algorithm to solve the optimization problem involved in this model.
- We demonstrated numerically the performance of our algorithm on both simulated and real datasets.
- We illustrated the benefit of such method to study electromagnetic signals form recorded from brain activity.

Thanks for your attention!

Code available online:

alphacsc : alphacsc.github.io

DiCoDiLe : github.com/tommoral/dicodile

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.

Buzsáki, 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.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

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 Le Cun, 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*, 1901.09235.

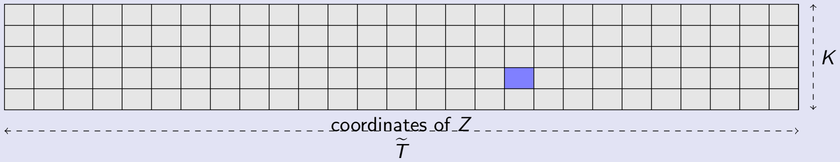
Moreau, T., Oudre, L., and Vayatis, N. (2018). DICOD: Distributed Convolutional Sparse Coding. In *International Conference on Machine Learning (ICML)*, pages 3626–3634, Stockohlm, Sweden. PMLR (80).

Nesterov, Y. (2010). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362.

Osher, S. and Li, Y. (2009). Coordinate descent optimization for ℓ_1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Problems and Imaging*, 3(3):487–503.

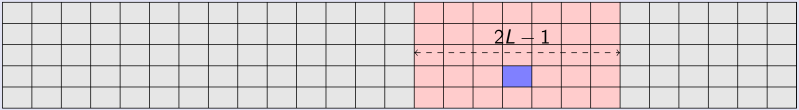
Wright, S. and Nocedal, J. (1999). *Numerical optimization*. Science Springer.

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



GCD has $\mathcal{O}(K\widetilde{T})$ computational complexity.

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

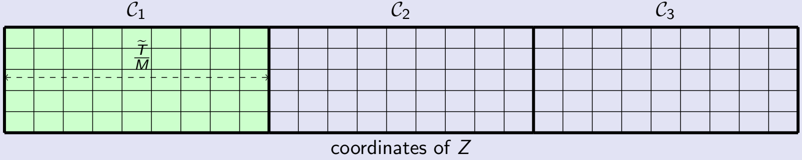


coordinates of Z

GCD has $\mathcal{O}(K\widetilde{T})$ computational complexity.

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

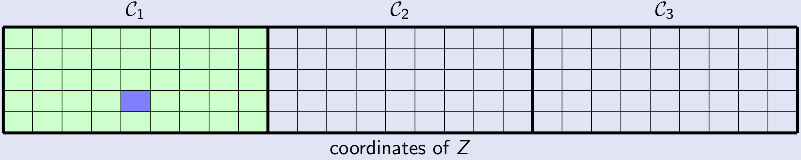
We introduced the LGCD method which is an extension of GCD.



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

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

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



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

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

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

$$\frac{\widetilde{T}}{M} = 2L - 1 \quad \Rightarrow \quad \mathcal{O}(\text{Coordinate selection}) = \mathcal{O}(\text{Coordinate Update})$$

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

$\Rightarrow$ Efficient for sparse Z

D-step: solving for the atoms

We use the projected gradient descent with an Armijo backtracking line-search Wright and Nocedal [1999] for both u-step and v-step for

$$\min_{\substack{\|u_k\|_2\leq 1\\ \|v_k\|_2\leq 1}} E(u_k, v_k) \triangleq \sum_{n=1}^{N} \frac{1}{2}\|X^n - \sum_{k=1}^{K} z_k^n * (u_k v_k^\top)\|_2^2 \quad . \tag{3}$$

One important computation trick is for fast computation of the gradient.

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

Computing $\nabla_{D_k} E(u_k, v_k)$ can be done efficiently

$$\nabla_{D_k} E(u_k, v_k) = \sum_{n=1}^{N} (z_k^n)^{\Lsh} * \left(X^n - \sum_{l=1}^{K} z_l^n * D_l \right) = \Phi_k - \sum_{l=1}^{K} \Psi_{k,l} * D_l \ ,$$

Pattern recovery

Test the pattern recovery capabilities of our method on simulated data,

$$X^n = \sum_{k=1}^{2} z_k * (u_k v_k^\top) + \mathcal{E}$$

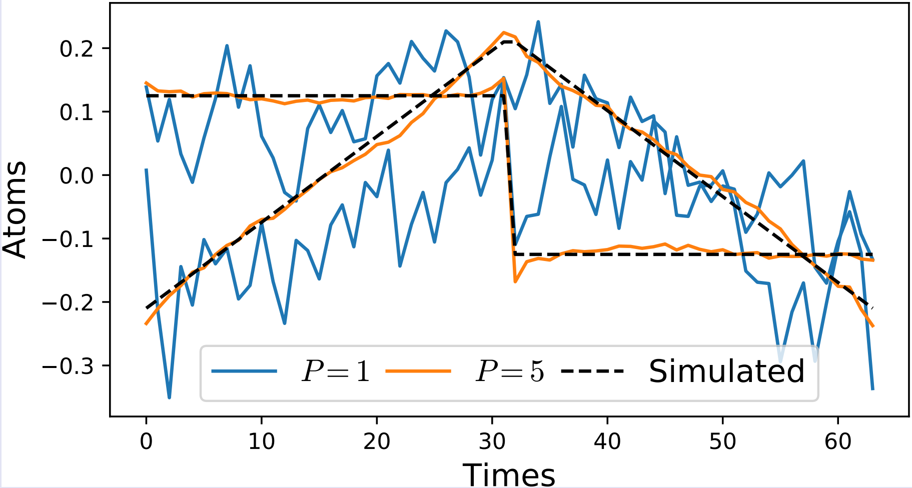
where (u_k, v_k) are chosen patterns of rank-1 and the activated coefficient $z_k^n[t]$ are drawn uniformly and their value are uniform in $[0, 1]$.

The noise $\mathcal{E}$ is generated as a gaussian white noise with variance σ.

We set $N = 100$, $L = 64$ and $\widetilde{T} = 640$

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.