Compressed Online Dictionary Learning for Fast Resting-State fMRI Decomposition

Arthur Mensch[1] Gaël Varoquaux[1] Bertrand Thirion[1]

[1]Parietal team, Inria, CEA, Paris-Saclay University. Neurospin, 91191 Gif-sur-Yvette, France

Sparse decomposition for resting-state fMRI

- Set of t records holding fMRI resting-state sequences (4D, n samples, p voxels)
- t matrices $(\mathbf{X}^s)_{s\in[1,t]}$ in $(\mathbb{R}^{n\times p})^t$

Find $k \ll n, p$ sparse spatial components in $\mathbb{R}^p$ such that :

$$\forall s \in [\![1, t]\!],\ \mathbf{X}^s = \mathbf{U}^s\mathbf{V}^\top \quad \text{with} \quad \mathbf{U}^s \in \mathbb{R}^{n\times k},\ \mathbf{V} \in \mathbb{R}^{p\times k}$$

Low-rank sparse decomposition methods for rfMRI

ICA + thresholding   Dictionary Learning

Dictionary learning outperforms ICA [3] → better **stability / explained variance tradeoff**

Dictionary learning

$$\min_{\substack{\mathbf{V}\in\mathbb{R}^{p\times k}\\ \mathbf{U}\in\mathbb{R}^{nt\times k}, \|\mathbf{U}_j\|_2\leq 1}} \|\mathbf{X} - \mathbf{U}\,\mathbf{V}^\top\|_F^2 + \lambda\,\|\mathbf{V}\|_1 \qquad (1)$$

- Record concatenation $\mathbf{X} = [\mathbf{X}^0, \dots, \mathbf{X}^{t-1}]^\top$
- ℓ_2 reconstruction loss
- Sparsity inducing ℓ_1 penalty
- k dense temporal atoms in $\mathbb{R}^{n\,t}$
- k loadings in $\mathbb{R}^p$: sparse spatial maps

Online DL

- Learn $\mathbf{U}$ streaming $\mathbf{X}^t \in \mathbb{R}^{n\,t}$ [2]
- Learn $\mathbf{V}$ solving a large Lasso problem
- Single iteration in $n\,t$
- Few passes : algorithm scales in $\mathcal{O}(n\,t \times p)$

Scaling DL to large-scale rfMRI data

- Online dictionary learning designed for small sample dimension $n\,t$
- **HCP dataset** : $n\,t \sim 4\cdot 10^6$, $p = 2\cdot 10^5$
- Prohibive single iteration cost $n\,t$
- Memory access in different directions → need to avoid out-of-core computation

Can we reduce n and still obtain good spatial maps $\mathbf{V} \in \mathbb{R}^{k\times p}$?

Sketched dictionary learning problem

- Good quality maps from small datasets (ADHD dataset, $n\,t \sim 7000$)
- Reduce $n \to m$ by dimension reduction
- Projection / subsampling : $\mathbf{P} \in \mathbb{R}^{mt\times nt}$

$$\min_{\substack{\mathbf{V}\in\mathbb{R}^{p\times k}\\ \mathbf{U}_r\in\mathbb{R}^{mt\times k}, \|(\mathbf{U}_r)_j\|_2\leq 1}} \left\|\mathbf{P}\,\mathbf{X} - \mathbf{U}_r\mathbf{V}^\top\right\|_F^2 + \lambda\,\|\mathbf{V}\|_1 \qquad (2)$$

$\mathbf{V} \in \mathbb{R}^{k\times p}$ as in (1) → valid spatial components

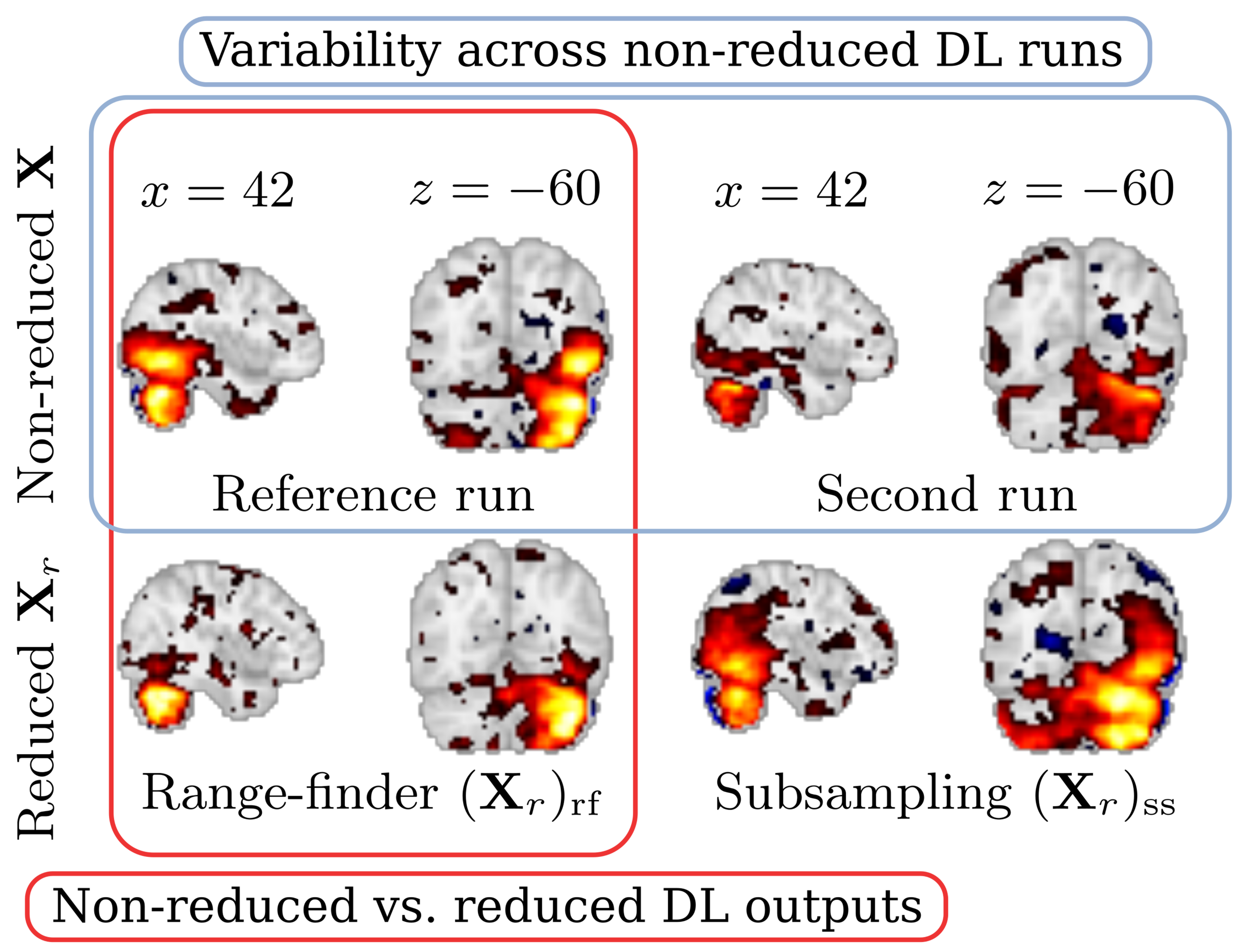

Figure 3: Maps obtained with dictionary learning + reduction are qualitatively equivalent to maps obtained with non-reduced DL

Dimension reduction

- **Per subject reduction** $\mathbf{X}_r^s \in \mathbb{R}^{m\times p} = \mathbf{P}^s\mathbf{X}^s$
- $\mathbf{P}$ block diagonal : $\mathbf{X}_r = [\mathbf{X}_r^0, \dots, \mathbf{X}_r^{t-1}]^\top$

Different sketches

- Optimal projector — SVD on $\mathbf{X}^s$ with prohibitive $\mathcal{O}(np^2)$ cost
- **Range-finding** random projection ($\mathcal{O}(npm)$) : $\mathbf{P}^s$ such that

$$\|\mathbf{X}^s - \mathbf{P}^s\,{\mathbf{P}^s}^\top\mathbf{X}^s\|_F \approx \min_{\mathbf{Y}^s\in\mathbb{R}^{n\times p}\text{rank}(\mathbf{Y}^s)\leq m} \|\mathbf{X}^s - \mathbf{Y}^s\|_F \qquad (3)$$

- Random subsampling : $\mathbf{P}^s \in \mathbb{R}^{m\times n}$ all zero but one per line

Validation

- Effect of dimension reduction $\mathbf{X}, \mathbf{PX} \to \mathbf{V}^0, \mathbf{V}$
- Compare $\mathbf{V}^0$ and $\mathbf{V}$: Hungarian algorithm + mean correlation

$$d(\mathbf{V}, \mathbf{V}^0) = \frac{1}{k}\max_{\boldsymbol{\Omega}\in\mathcal{S}_k} \mathrm{Tr}\left(\mathbf{V}^\top\boldsymbol{\Omega}\,\mathbf{V}^0\right) \in [0, 1] \qquad (4)$$

⚠ Stochastic algorithm → random output

- Stack 10 runs for each reduction method → $\mathbf{V}^{\text{aug}}$
- Mitigate stochastic effect
- **Reference** : 10 runs of vanilla algorithm (*pseudo ground-truth*)
- Compare random projection and random subsampling

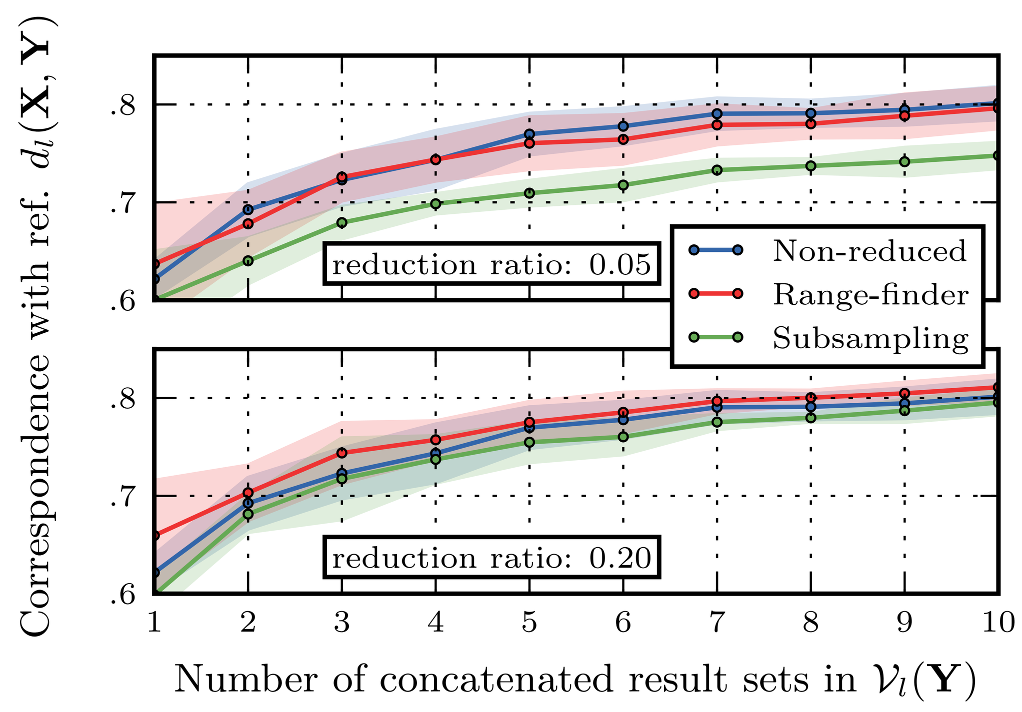

Figure 1: DL + RF is as accurate as vanilla DL when stacking output

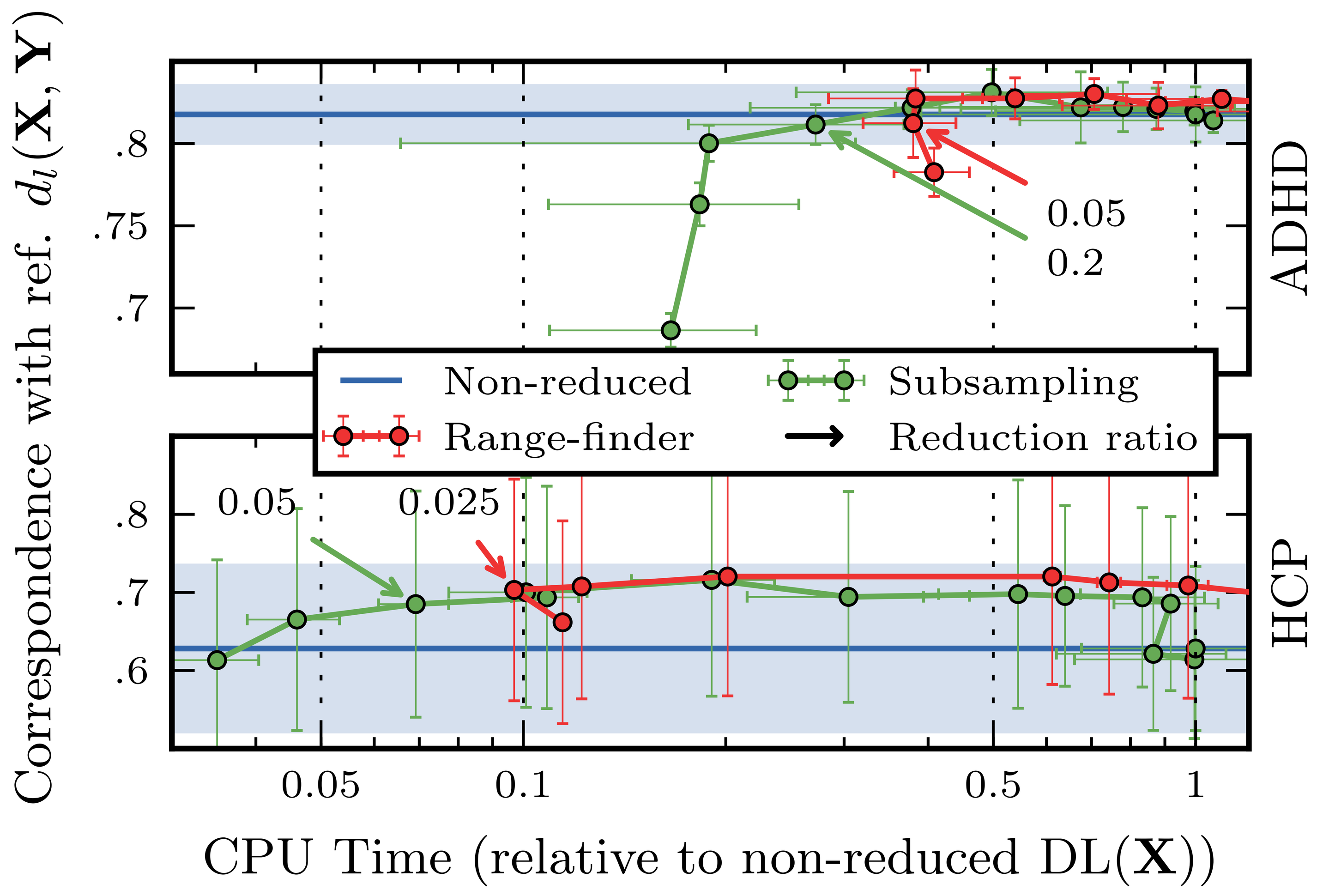

Figure 2: $\mathbf{V}$ accuracy is stable as: ↑ reduction ratio $\frac{n}{m}$ ↓ CPU time

Results

- Online DL for resting-state fMRI is **robust to dimension reduction** in the time direction $n \to m$
- **Range-finding performs better** than random subsampling on two different datasets
- With very little overhead compared to DL runtime

		CPU Time		Correspondence $d_l(\mathbf{X}, \mathbf{Y})$	
Data	Reduc.	Red.	N-red.	Reduced	Non-reduced
HCP	.025	**849 s**	7425 s	**.703** ± **.141**	.628 ± .105
ADHD	.05	**71 s**	186 s	**.796** ± **.020**	.801 ± .016

Table: Accuracy and DL speed-up using range-finder projectors

In practice

On a 75 record subset of HCP (**150GB**)

- **10x speed-up** in time + memory usage
- **Non-significant accuracy loss**

Please use it !

```python
from nilearn.datasets import fetch_adhd
from nilearn.decomposition import DictLearning
from nilearn.plotting import plot_prob_atlas, show

dataset = fetch_adhd(n_subjects=40).func
decomposer = DictLearning(n_components=30, alpha=1, reduction=5)
decomposer.fit(dataset)
masker = decomposer.masker_
components = masker.inverse_transform(decomposer.components_)
plot_prob_atlas(components); show()
```

[1] A. Mensch, G. Varoquaux, and B. Thirion. Compressed online dictionary learning for fast fmri decomposition. *IEEE International Symposium on Biomedical Imaging*, Prague, CZ, Apr. 2016.

[2] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[3] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information Processing in Medical Imaging*, pages 562–573. Springer, 2011.