Online Dictionary Learning with Group Structure Inducing Norms

Zoltán Szabó[1] **Barnabás Póczos**[2] **András Lőrincz**[1]

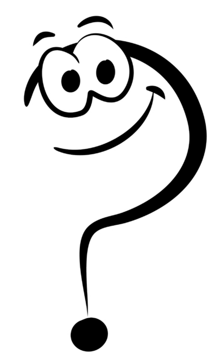
[1]Faculty of Informatics, Eötvös Loránd University
Pázmány Péter sétány 1/C, Budapest, H-1117, Hungary
email: `{szzoli,andras.lorincz}@elte.hu`
web: `http://nipg.inf.elte.hu`

[2]School of Computer Science, Carnegie Mellon University
5000 Forbes Ave, 15213, Pittsburgh, PA, USA
email: `bapoczos@cs.cmu.edu`
web: `http://www.autonlab.org`

1. Introduction

- Sparse coding.
- Structured sparsity (e.g., disjunct groups, trees): increased performance in several applications.
- **Our goal**: develop a dictionary learning method, which
  - enables general overlapping group structures,
  - is online: fast, memory efficient, adaptive,
  - applies non-convex sparsity inducing regularization:
    - fewer measurements,
    - weaker conditions on the dictionary,
    - robust (w.r.t. noise, compressibility).
  - can deal with missing information.

Current approaches can exhibit two of these features at most.



2. Problem

Task:

- Group structure inducing on the hidden representation $\boldsymbol{\alpha}$ through regularization:

$$\Omega(\boldsymbol{\alpha}) = \|(\|\boldsymbol{\alpha}_G\|_2)_{G\in\mathcal{G}}\|_\eta, \tag{1}$$
$$\Omega(\boldsymbol{\alpha}) = \|(\|\mathbf{d}^G \circ \boldsymbol{\alpha}\|_2)_{G\in\mathcal{G}}\|_\eta, \tag{2}$$
$$\Omega(\boldsymbol{\alpha}) = \|(\|\mathbf{A}^G \boldsymbol{\alpha}\|_2)_{G\in\mathcal{G}}\|_\eta, \quad \eta \in (0,2). \tag{3}$$

- Approximate on the observed coordinates ($\mathbf{x}_O$) using dictionary $\mathbf{D}$:

$$\frac{1}{2}\|\mathbf{x}_O - \mathbf{D}_O\boldsymbol{\alpha}\|_2^2. \tag{4}$$

- Loss for a fixed observation ($\kappa > 0$):

$$l(\mathbf{x}_O, \mathbf{D}_O) = \min_{\boldsymbol{\alpha}}\left[\frac{1}{2}\|\mathbf{x}_O - \mathbf{D}_O\boldsymbol{\alpha}\|_2^2 + \kappa\Omega(\boldsymbol{\alpha})\right]. \tag{5}$$

- Goal: minimize online the average loss of the dictionary

$$\min_{\mathbf{D}} f_t(\mathbf{D}) := \frac{1}{t}\sum_{i=1}^{t} l(\mathbf{x}_{O_i}, \mathbf{D}_{O_i}). \tag{6}$$

- Possible dictionary/representation constraints:
  - $\mathbf{D} \in \mathcal{D} = \times_{i=1}^{d_\alpha}\mathcal{D}_i \subseteq \mathbb{R}^{d_x}$: closed, convex, and bounded.
  - $\boldsymbol{\alpha} \in \mathcal{A} \subseteq \mathbb{R}^{d_\alpha}$: convex, closed.

3. Special cases

$O_i = \{1,\ldots,d_x\}$ $(\forall i)$: fully observed OSDL task.

Special cases for $\mathcal{G}$:

'Traditional' sparse dictionary	$\mathcal{G}$ = $\{\{1\},\{2\},\ldots,\{d_\alpha\}\}$.
Hierarchical dictionary	$\mathcal{G}$ = descendants of the nodes.
Grid adopted dictionary	$\mathcal{G}$ = nearest neighbors of the nodes.
Group Lasso	$\mathcal{G}$ = partition.
Elastic net	$\mathcal{G}$ = singletons and $\{1,\ldots,d_\alpha\}$.
Contiguous, nonzero representations	$\mathcal{G}$ = intervals.

Special cases for $\mathcal{D},\mathcal{A}$:

'Traditional' setting	ℓ_2 constrained $\mathbf{D}$.
Structured NMF	non-negative $\mathbf{D}$ and $\boldsymbol{\alpha}$.
Structured mixture-of-topics	ℓ_1 constained $\mathbf{D}$, non-negative $\mathbf{D}$, $\boldsymbol{\alpha}$.
'Hard' representation constraints	group norm/elastic net/fused Lasso constrained $\boldsymbol{\alpha}$.
Double structured dictionaries	group norm constraints to $\boldsymbol{\alpha}$ *and* $\mathbf{D}$.

Special cases for $\{\mathbf{A}^G\}_{G\in\mathcal{G}}$:

Fused Lasso	$\Omega(\boldsymbol{\alpha})$ = $\sum_{j=1}^{d_\alpha-1}\|\alpha_{j+1}-\alpha_j\|$
Graph-guided fusion penalty	$\Omega(\boldsymbol{\alpha})$ = $\sum_{e=(i,j)\in E:i<j} w_{ij}\|\alpha_i - v_{ij}\alpha_j\|$
Linear trend/polynomial filtering	$\Omega(\boldsymbol{\alpha})$ = $\sum_{j=2}^{d_\alpha-1}\|-\alpha_{j-1}+2\alpha_j-\alpha_{j+1}\|$
Generalized Lasso penalty	$\Omega(\boldsymbol{\alpha})$ = $\\|\mathbf{A}\boldsymbol{\alpha}\\|_1$
Total variation	$\Omega(\boldsymbol{\alpha})$ = $\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\\|(\nabla\boldsymbol{\alpha})_{ij}\\|_2$

4. Optimization

Online optimization of dictionary $\mathbf{D}$ through alternations:

1. $(\mathbf{x}_{O_t}, \mathbf{D}_{t-1}, O_t) \mapsto \boldsymbol{\alpha}_t$:

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha}\in\mathcal{A}}{\operatorname{argmin}}\left[\frac{1}{2}\|\mathbf{x}_{O_t} - (\mathbf{D}_{t-1})_{O_t}\boldsymbol{\alpha}\|_2^2 + \kappa\Omega(\boldsymbol{\alpha})\right]. \tag{7}$$

Solution idea: iterated reweighted least squares using the variational property of $\|\cdot\|_\eta$.

2. $\{\boldsymbol{\alpha}_i\}_{i=1}^t \mapsto \mathbf{D}_t$ by means of quadratic optimization:

$$\hat{f}_t(\mathbf{D}_t) = \min_{\mathbf{D}\in\mathcal{D}} f_t(\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^t). \tag{8}$$

Solution idea:

- block-coordinate descent optimization: update column $\mathbf{d}_j$, while keeping the others fixed,
- statistics of the cost $\hat{f}_t$ can be efficiently updated online (matrix recursions).

5. Numerical experiments

5.1 Inpainting of natural images

We focused on the following questions:

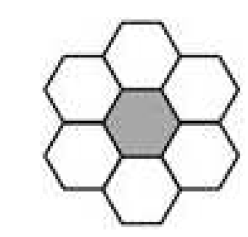
- structured (toroid) vs. unstructured dictionary for inpainting,
- efficiency in case of missing observations,
- inpainting of *full images* using dictionaries learned on partially observed patches.

First experiment (complete observation):

- increasing neighbor size = decreasing MSE.
- $r = 3$: $13-19\%$ improvement compared to the unstructured case ($r = 0$).

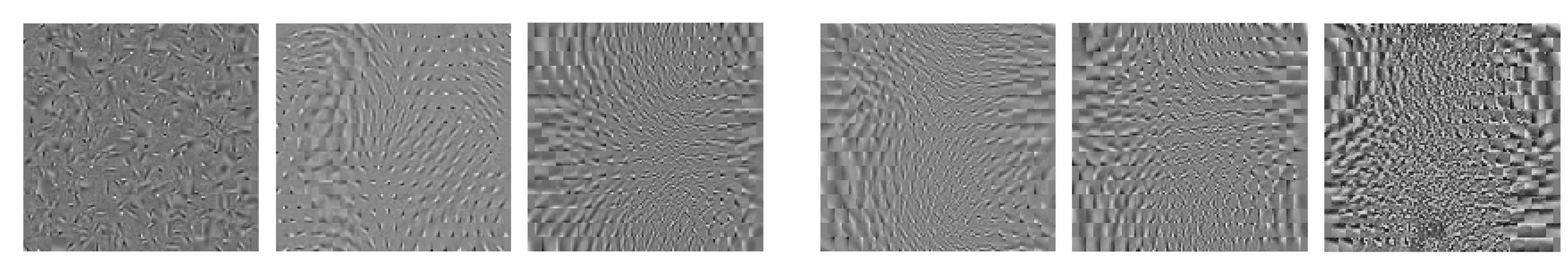
Second experiment (neighbor size: $r = 3$, missing pixels: $p_{tr} \le 0.9$):

- Up to about $p_{tr} = 0.7$: MSE grows slowly.
- $\mathbf{D}$-s in Fig. 1(d)-(f).
- For $p_{tr} = 0.9$, MSE still relatively small, see Fig. 2(a).



Third experiment (neighbor size: $r = 3$, missing pixels: $p_{tr} = 0.5$):

- Task: inpainting of a *full* unseen image.
- Result: sliding average, Fig. 2(a), $p_{test}^{val} = 0.7$, PSNR = 29 dB.

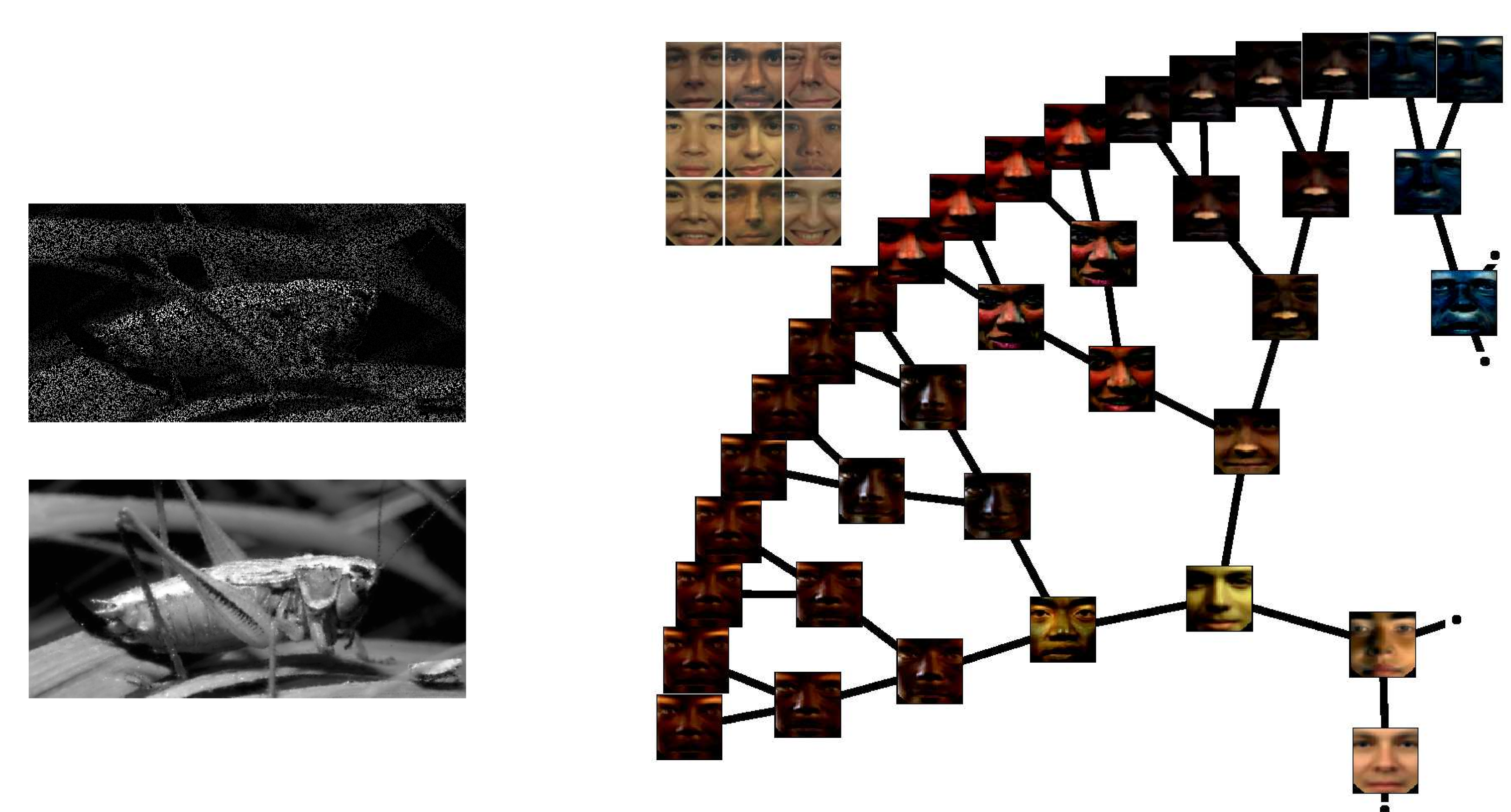


(a) (b) (c) (d) (e) (f)

Figure 1: *Group-structured* $\mathbf{D}$*-s. (a)-(c): complete; increasing neighbor size* ($r = 0, 2, 3$)*. (d)-(f): increasing incompleteness* ($p_{tr} = 0, 0.1, 0.5$)*.*

5.2 Online structured non-negative matrix factorization on faces

- Online, $\mathcal{G}$-NMF: special case of OSDL.
- Illustration: color FERET large-scale (140×120) facial dataset.
- $\mathcal{G}$: complete, 8-level binary tree ($d_\alpha = 255$).



(a) (b)

Figure 2: *(a): full image inpainting illustration; top: observed, bottom: estimated. (b): structured NMF dictionary, training samples at the upper left corner.*

5.3 Collaborative Filtering

- Joke recommendation (Jester): 100 jokes $\times$ $73,421$ users.
- Observation: $\mathbf{x}_{O_t}$ = ratings of the t^{th} user.
- Baseline: best known RMSE = 4.1123 (item neighbor), 4.1229 (unstructured dictionary, $d_\alpha = 100$).
- Result: toroid $\mathcal{G}$ ($d_\alpha = 100$) $\rightarrow$ RMSE = 4.0774, hierarchical ($d_\alpha = 15$) $\rightarrow$ 4.1220.

The research was partly supported by the Department of Energy (grant number DESC0002607).

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638

MAGYARORSZÁG MEGÚJUL

The Project is supported by the European Union and co-financed by the European Social Fund (grant agreements no. TÁMOP 4.2.1/B-09/1/KMR-2010-0003 and KMOP-1.1.2-08/1-2008-0002).