Signals, images and task-driven representation and dictionary learning

November 24, 2016

Keywords: dictionary learning, supervised dictionary learning, transform learning, sparse recovery, classification, feature engineering

1 Context

We live in an era of the data abundance: tremendous volume of data are available in regards to virtually everything. It includes biomedical signals, genomic data, industrial systems signals, weather and environment related data, astronomical surveys data, satellite earth images, and a plethora of human activities generated data. Also, spectacular breakthroughs in data analysis techniques and methodology have occurred in signal processing and machine learning communities over the two past decades. One can get interesting insights on the convergence between these two areas from [Singh et al. 2015]. At the intersection of these two fields, Dictionary Learning (DL) has come to be a major topic of research on itself. It encompasses methods and algorithms which, for a given type of signal, aim at deriving a set of cardinal features which enables one to concisely describe signals of this type.

Precisely, given a set of vectors \( (x_j)_{1 \leq j \leq n} \) in \( \mathbb{R}^p \) that might represent ECG signals or genes sequences as illustrated in Fig.1 DL techniques compute a set vectors \( (d_i)_{1 \leq i \leq N} \) and a set of coefficients \( (\alpha_{ij})_{1 \leq i \leq N, 1 \leq j \leq n} \) so that

\[
x_j \approx \sum_{i=1}^{N} \alpha_{ij} d_i, \quad j = 1 \cdots N
\]

with a constraint of sparsity on the vectors \( \alpha_j = [\alpha_{1j}, \cdots, \alpha_{Nj}]^T \). As for the terminology,

- the matrix \( D = [d_1, \cdots, d_N] \) is the dictionary,
• its columns vectors \((d_i)\) are the atoms,

• and the vectors \((\alpha_j)\) are termed the codes as they represent each of the training set samples.

Hence, the learnt dictionary is specifically adapted for describing signals that are structurally similar to the training samples. The benefit of such dictionaries in sparsity-driven signal recovery has been shown in several applications (see for example \cite{Elad2006, Ravishankar2011, Beckouche2013}).

10 years after the seminal contribution in \cite{Aharon2006}, major advances have been realized in DL regarding different important aspects including:

• discriminatory power: in numerous applications in machine learning, data are labeled and/or sampled from some regular manifold; thus it is suitable, for classification or interpolation tasks for instance, that the learned codes allow for a better discrimination of the data samples with respect to labels information or manifold’s structure; the growing field of supervised dictionary learning precisely consists of DL methods that account for these additional information (a recent review can be found in \cite{Gangeh2015});

• geometric invariances: a considerable effort has been dedicated into to building DL methods that use the fact that similar structures can appear in the training set with different offsets, orientations (for bidimensional data and above) and scales; such methods rely either on training data registration \cite{Beckouche2013, Yuzuguler2014} or on learning structured dictionaries \cite{Barthelemy2012, Ophir2011};

• scalability: generally speaking, the main disadvantage of learned dictionaries with respect to analytical dictionaries such as wavelets, DCT or curvelets related dictionaries is that the latter can be used through fast implicit transforms while the former are unstructured and potentially large matrices that one has to manipulate explicitly, which limits the practioner to learning reasonably small dictionaries on low dimensional data; this issue has been addressed in several works either from the data dimensionality angle \cite{Xiang2011, Sulam2016} or from the computational efficiency of the dictionary perspective \cite{Magoarou2014, Chabiron2014}.

Such rich perspectives make DL an exciting and promising field of research, as far as data analysis is concerned; this Ph.D. thesis is precisely positioned at the confluence of these three sub-themes.

2 Research project

The goal of this thesis is to propose a dictionary and representation learning methodology that yields

• a good tradeoff between data description and discrimination

• and easily interpretable dictionary and codes.

As for the first point, the aim will be to take a step further in supervised dictionary learning by introducing geometric invariances. The possibility of learning a "fast transform like" dictionary in this context will be examined and ideas from the transform learning literature will be considered (see for instance \cite{Pfister2015} and the references therein).

Supervised DL methods in general have important hyperparameters which choice is not straightforward and which are costly to tune. Proposing a rigorous framework for setting learning hyperparameters constitutes an important aspect of this thesis and will rely on existing solutions proposed in the machine learning and signal processing communities (see for example \cite{Pedregosa2016, Chaux2012, Dang2015, Shervashidze2015} ).

An interesting line of investigation would be to evaluate the feasibility and benefit of adopting a non-linear descriptive model in place of the model in Eq.1 in the same spirit as kernel-based DL method for instance (see \cite{Golts2016} and the references therein).
The second point is application oriented. Indeed, it is suitable to produce physically meaningful or interpretable atoms and codes in DL, especially when analyzing physiological signals. This notion of interpretability is both ubiquitous and ill-defined in machine learning as pointed out in [Lipton et al. 2016]. The reflection here will be on whether it is possible to define a quantitative interpretability criterion and use it in the learning procedure. Such criterion might be generic as the sharpness index introduced in [Leclaire & Moisan 2013] or the typical criteria used for points of interest detection in images (see for instance [Lowe 2004]). Alternately, it might rely on application specific quantities such as the Mel Frequency Cepstral Coefficients [Muda et al. 2010] in voice recognition.

A broad spectrum of applications can be considered including common classification tasks (handwritten digits, natural images...), biophysical signals analysis, hyperspectral earth images analysis, galaxies images or spectra classification, astronomical radio sources identification etc.

3 Organisation of the thesis work

The thesis will be organised in three parts:

1. Supervised DL
   - State-of-the-art: Supervised DL, DL with geometric invariance, fast transform like DL
   - Methodology: supervised DL+invariances
   - Methodology: supervised DL+invariances+computational efficiency

2. Hyperparameters
   - State-of-the-art: hyperparameters selection and optimization in machine learning and signal processing
   - Methodology: supervised DL+invariances+computational efficiency+automated parameters selection

3. Interpretability
   The programming will be done in Python.

4 The candidate

- **Education**: Master 2 or equivalent in Signal Processing or Applied Mathematics or Data Science
- **Skills**: knowledge and practice of Optimization numerical schemes is required; an experience in Python programming is a plus
- **Personal**: autonomous, quick learner

5 Why apply?

5.1 The environment

This Ph.D. thesis will proceed in co-tutoring between the [LADIS] and the [CosmoStat] laboratories at CEA Saclay’s Technological and Fundamental Research Directions respectively. These young and dynamic laboratories have a recognized expertise

- in data analysis methodology and industrial applications for the LADIS
- and in signal processing methodology and applications to astronomy for the CosmoStat.

This double culture will provide the Ph.D. candidate a wealthy and structuring learning ground and a high visibility as for professional openings.
5.2 Learnings

The Ph.D. candidate will master modern and advanced data analysis tools, at the intersection of signal processing and machine learning. This thesis will provide the future Doctor a solid base for an orientation toward both academical research and research engineering in data science.

- **contacts:** Jean-Luc Starck (Ph.D. supervisor), jean-luc.starck@cea.fr; Ngolé Fred (tutor), fred-maurice.ngole-mboula@cea.fr;
- **Lab:** CEA/LADIS-CosmoStat in Saclay.

References


