Learning Diverse Overcomplete Dictionaries via Determinantal Priors

Maruan Al-Shedivat ALSHEDIVAT@CS.CMU.EDU
Yo Joong Choe YJCHOE@CMU.EDU
Neil Spencer NSPENCER@ANDREW.CMU.EDU
Eric P. Xing EPXING@CS.CMU.EDU
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

1. Motivation

High-dimensional signals can often be represented as linear combinations of elementary functions called *atoms* from a collection called the *dictionary*, which can be either designed or learned from the data. An *overcomplete* dictionary, which consists of many similar atoms that are only sparsely used, can allow for accurate and robust representations of the signals (Lewicki & Sejnowski, 2000).

The predominant guiding principle of dictionary design or learning has been to encourage sparsity in the eventual representation of the data. This principle reduces the problem of signal decomposition to an ℓ_0- or ℓ_1-norm optimization problem. While this approach is intuitive, in order to be consistent, it requires strong assumptions on the nature of the signals, such as *mutual incoherence* (Donoho & Huo, 2001) or *restricted isometry* (Candes & Tao, 2005) of the overcomplete dictionary.

To tackle this issue, we propose a model that additionally encourages *diversity* of the signal representation. The diversity of the learned dictionary atoms is induced by using the *determinantal point process* (Kulesza & Taskar, 2012) as a prior in the probabilistic formulation of the dictionary learning problem. This approach is useful because it only requires incoherence between each small set of atoms chosen to represent each signal instead of the global incoherence of the entire dictionary. Moreover, the probabilistic formulation allows to view the model from a geometric perspective, as it defines a distribution over the Grassmannian.

2. Geometric Perspective

Representing data with an overcomplete dictionary results in assigning different points to different subsets of atoms, or equivalently to different linear or affine subspaces of the ambient space. In this sense, dictionary learning seeks to find a collection of potentially low-rank subspaces that represent the data effectively. Hence, it is closely connected to the problem of subspace segmentation (Liu et al., 2010).

From the geometric perspective, dictionary learning identifies a collection of points on the Grassmann manifold that correspond to the subspaces that contain the signal. Our probabilistic model defines a random process on the Grassmann manifold, where the regions of high posterior probability are likely to be the subspaces that contain the data. We believe that casting the problem into the geometric framework can help to formalize the intuitions behind the diverse dictionaries and provably guarantee consistency of the model under mild conditions such as incoherence of only the sub-dictionaries that represent the signal.

3. Model and Discussion

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ be a collection of data. Given a dictionary $\boldsymbol{D} \in \mathbb{R}^{p \times m}$, we consider the following data-generating process:

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{DPP}(\boldsymbol{D}^\top \boldsymbol{D}), \\ \boldsymbol{\xi}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi}), \\ \boldsymbol{\varepsilon}_i &\sim \mathcal{N}(0, \sigma^2 I_p), \\ \boldsymbol{x}_i &= \boldsymbol{D}(\boldsymbol{z}_i \odot \boldsymbol{\xi}_i) + \boldsymbol{\varepsilon}_i, \end{aligned}$$

where $\boldsymbol{\varepsilon}_i$ is an unknown noise vector, $\boldsymbol{z}_i \in \{0,1\}^m$ is an indicator-vector that selects a sub-dictionary, $\odot$ denotes the element-wise multiplication, $\mathcal{DPP}(\boldsymbol{D}^\top \boldsymbol{D})$ is the deteminantal point process (DPP) with the kernel defined via the dot product between the dictionary atoms. In this scenario, DPP assigns probability to the subsets of atoms as follows:

$$\mathbb{P}(\boldsymbol{z}) = \frac{\det(\boldsymbol{D}_{[\boldsymbol{z}]}^\top \boldsymbol{D}_{[\boldsymbol{z}]})}{\det(\boldsymbol{D}^\top \boldsymbol{D} + I_m)}.$$

Given the probabilistic model, one seeks a dictionary with high posterior probability $\mathbb{P}(\boldsymbol{D} | \boldsymbol{X})$. Currently, our model is trained heuristically using a Monte-Carlo EM algorithm and is able to recover the true overcomplete dictionaries consistently, even when the dictionaries are *not* incoherent. Ultimately, our goal is to cast the problem into the geometric framework to derive more tractable and provably accurate inference strategies.

References

Candes, Emmanuel J and Tao, Terence. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

Donoho, David L and Huo, Xiaoming. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2001.

Kulesza, Alex and Taskar, Ben. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3), 2012.

Lewicki, Michael S and Sejnowski, Terrence J. Learning overcomplete representations. *Neural computation*, 12 (2):337–365, 2000.

Liu, G, Lin, Z, and Yu, Y. Robust subspace segmentation by low-rank representation. In *ICML*, pp. 663–670, 2010.