Proximal Methods for Sparse Hierarchical Dictionary Learning

Rodolphe Jenatton[1] rodolphe.jenatton@inria.fr
Julien Mairal[1] julien.mairal@inria.fr
Guillaume Obozinski[1] guillaume.obozinski@inria.fr
Francis Bach[1] francis.bach@inria.fr

[1] *INRIA - WILLOW Project-team,*
Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548),
23, avenue d'Italie, 75214 Paris. France

February 2, 2010

Abstract

This paper proposes to combine two approaches for modeling data admitting sparse representations: On the one hand, dictionary learning has proven very effective for various signal restoration and representation tasks. On the other hand, recent work on structured sparsity provides a natural framework for modeling dependencies between dictionary elements. We propose to combine these approaches to learn dictionaries embedded in a hierarchy. We show that the proximal operator for the tree-structured sparse regularization that we consider can be computed exactly in linear time with a primal-dual approach, allowing the use of accelerated gradient methods. Experiments show that for natural image patches, learned dictionary elements organize themselves naturally in such a hierarchical structure, leading to an improved performance for restoration tasks. When applied to text documents, our method learns hierarchies of topics, thus providing a competitive alternative to probabilistic topic models.

Learned sparse representations, initially introduced by Olshausen and Field [1997], have been the focus of much research in machine learning, signal processing and neuroscience, leading to state-of-the-art algorithms for several problems in image processing. Modeling signals as a linear combination of a few "basis" vectors offers more flexibility than decompositions based on principal component analysis and its variants. Indeed, sparsity allows the use of overcomplete dictionaries, which have a number of basis vectors greater than the original signal dimension.

As far as we know, while much attention has been given to efficiently solving the corresponding optimization problem [Mairal et al., 2009], there are few attempts in the literature to make the model richer by adding structure between dictionary elements. We propose to use recent work on structured sparsity [Zhao et al., 2009, Kim and Xing, 2009, Jenatton et al., 2009] to embed the dictionary elements in a hierarchy.

Hierarchies of latent variables, typically used in neural networks and deep learning architectures have emerged as a natural structure in several applications, notably to model text documents. Indeed, in the

context of *topic models*, hierarchical models using Bayesian non-parametric methods have been proposed by Blei et al. [2010]. Quite recently hierarchies have also been considered in the context of kernel methods [Bach, 2009], where they are motivated by computational reasons. To the best of our knowledge, although structured sparsity has already been used for regularizing dictionary elements by Jenatton et al. [2009], it has never been used to model dependencies between them. This paper makes three contributions:

• We propose to use a structured sparse regularization to learn a dictionary embedded in a tree.

• We show that the proximal operator for a tree-structured sparse regularization can be computed exactly in a finite number of operations using a primal-dual approach, with a complexity linear in the number of variables. This allows to use accelerated gradient methods [Beck and Teboulle, 2009] to solve tree-structured sparse decomposition problems, which may be useful in other uses of tree-structured norms Kim and Xing [2009], Bach [2009].

• Our method establishes a bridge between *dictionary learning for sparse coding* and *hierarchical topic models* [Blei et al., 2010]—to be viewed in the context of the correspondence between topic models and multinomial PCA [Buntine, 2002]—and can learn similar hierarchies of topics.

Our experiments show that for some classes of signals, learned dictionaries can benefit from such architectures, leading to similar or better restoration performance than those obtained by classical sparse coding, and that such a gain is also observed when our approach is used to learn topics in text documents.

References

F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, arXiv:0909.0844, 2009.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 2010. To appear.

W.L. Buntine. Variational Extensions to EM and Multinomial PCA. In *Proceedings of the 13th European Conference on Machine Learning*, page 34. Springer-Verlag, 2002.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009.

S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. Technical report, arXiv:0909.1373, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 2009. to appear, preprint: arXiv:0908.0050.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.