Structured sparse coding with a quadratic or bilinear penalty

Karol Gregor, Arthur Szlam, and Yann LeCun
NYU,
715 Broadway,
New York, NY,
kgregor@courant.nyu.edu, aszlam@courant.nyu.edu,
yann@courant.nyu.edu

February 10, 2011

Structured sparsity can be though of as a set of interactions between coefficients that modulate which dictionary atoms easily activate together. For example in the case of a tree, atoms on a given branch turn on together easily but atoms on different branches do not. In this work we start by introducing this interaction as a simple constraint between each pair of atoms in the dictionary; atoms can be simultaneously active only if there is no constraint between them. Then we will relax this constraint to a simple penalty. The resulting model is extremely flexible, and has the property that we can learn the structure penalty matrix S from data.

Suppose U is a set of disallowed index pairs $U = \{(i_1, j_1), (i_2, j_2), ..., (i_k j_k)\}$ for a representation Z of data points X via a dictionary W; here we will constrain $Z \geq 0$. The inference problem can be formulated as

$$\min_{Z \geq 0} \sum_{j=1}^{N} ||WZ_j - X_j||^2,$$

subject to

$$ZZ^T(i,j) = 0, \ i,j \in I.$$

Then the Langrangian of the energy with respect to Z is

$$\sum_{j=1}^{N} ||WZ_j - X_j||^2 + Z_j^T S Z_j, \tag{1}$$

where S_{ij} are the dual variables to each of the constraints in U, and are 0 in the unconstrained pairs. A local minimum of the constrained problem is a saddle point for (1). At such a point, S_{ij} can be interpreted as the weight

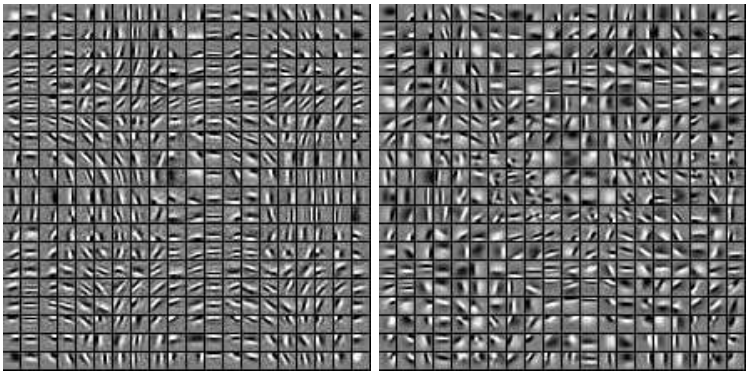

Figure 1: Filters learned from natural images in the setting where $S_{ij} = S^0 > 0$ if the distance between units i and j (arranged on a 2-d grid) satisfies $r_1 < d(i,j) < r_2$ and $S_{ij} = 0$ otherwise. (a) Images were preprocessed by local subtraction and contrast normalization with narrow width 1.65 pixels and trained in locally connected framework. The resulting edges are sharp and therefore can be naturally imbedded in two dimensions. (b) Same as (a) but only contrast subtraction was used with the width 5 pixels. Still similar edges are placed close together but because of the larger diversity of filters the pinwheels are not as smooth.

of the inhibitory connection between W_i and W_j necessary to keep them from simultaneously activating.

It can be useful to soften the constraints in U to a fixed, prespecified penalty, instead of a maximization over S as would be suggested by the Lagrangian form. This allows some points to use proscribed activations if they are especially important for the reconstruction. To use units with both positive and negative activations we take absolute values and obtain

$$\min_{W,Z} \sum_{x \in X} ||Wz - x||^2 + |z|^T S|z|, \tag{2}$$

$$||W_j|| = 1 \ \forall j.$$

Note that S will usually be chosen to be symmetric and have 0 on the diagonal.

In this work, we explore the model given by (2) and some variations, and show empirically that simple adaptations of standard sparse coding algorithms can be used for the inference and dictionary learning problems. We give examples of learning dictionaries arranged on trees, on 2-d topologies, and of learning the S matrix.