Supplementary Materials for Block and Group Regularized Sparse Modeling for Dictionary Learning

Yu-Tseh Chi* ychi@cise.ufl.edu
Mohsen Ali* moali@cise.ufl.edu
Ajit Rajwade† ajit_rajwade@daiict.ac.in
Jeffrey Ho* jho@cise.ufl.edu

Abstract

This document is the supplementary materials for the CVPR 2013 paper "Group and Sparse Regularized Sparse Modeling for Dictionary Learning". The first section contains derivations of the theoretical proof and of the two optimization algorithms in the Sec. 2 of the main text. The second section contains extra results from the experiment in Sec 3.2 of the main text.

1. Supplementary Materials to the Method Section

1.1. Simple Proof of the Theoretical Guarantee

Without loss of generality, we will prove the condition of one group data $\mathbf{X}$. We first concatenate columns of $\mathbf{X} \in \mathbb{R}^{n\times s}$ into a vector $\mathbf{x}' \in \mathbb{R}^{(n\cdot s)}$. The dictionary $\mathbf{D} = [\mathbf{d}_i \cdots \mathbf{d}_m]$ of m columns has to be converted accordingly to

$$\begin{aligned}\mathbf{D}' &= \left[\begin{array}{cccc|cccc} \mathbf{d}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{d}_2 & \mathbf{0} & \cdots & \cdots \mathbf{0} \\ \mathbf{0} & \mathbf{d}_1 & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{d}_2 & \cdots & \cdots \mathbf{0} \\ & & \vdots & & & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{d}_1 & \mathbf{0} & \mathbf{0} & \cdots & \cdots \mathbf{d}_m \end{array}\right] \\ &= [\mathbf{I}_s \otimes \mathbf{d}_1, \mathbf{I}_s \otimes \mathbf{d}_2 \cdots \mathbf{I}_s \otimes \mathbf{d}_m],\end{aligned}$$

where $\otimes$ denotes Kronecker product, and $\mathbf{I}_s$ is a $s \times s$ identity matrix. A block of k atoms, $\mathbf{d}_i \cdots \mathbf{d}_{i+k-1}$, in $\mathbf{D}$ is now a block of size $k \times s$ in $\mathbf{D}'$. We also concatenate rows of $\mathbf{C}$ into a vector $\mathbf{c}'$ as follows:

$$\mathbf{c}' = [\mathbf{C}_1\ \mathbf{C}_2 \cdots \mathbf{C}_m]^\intercal,$$

where $\mathbf{C}_i$ is the i-th row in $\mathbf{C}$. Assuming the number of atoms of the i-th block of the original dictionary $\mathbf{D}$ is n_{di}, then $\mathbf{C}_{[i]}$, the i-th block of coefficients $\mathbf{C}$, is now a segment of size $s \cdot n_{di}$ in $\mathbf{c}$. Our program

$$\mathcal{P}_{\ell_{0,p}} : \min_{\mathbf{C}} \sum_i I(\|\mathbf{C}_{[i]}\|_p) \ \text{ s. t. } \ \mathbf{X} = \mathbf{DC} \tag{1}$$

is now

$$\min_{\mathbf{c}'} \sum_i I(\|\mathbf{c}'_{[i]}\|_p) \ \text{ s. t. } \ \mathbf{x}' = \mathbf{D}'\mathbf{c}'. \tag{2}$$

Likewise,

$$\mathcal{P}_{\ell_{1,p}} : \min_{\mathbf{C}} \sum_i \|\mathbf{C}_{[i]}\|_p \ \text{ s. t. } \ \mathbf{X} = \mathbf{DC}. \tag{3}$$

is now

$$\min_{\mathbf{c}'} \sum_i \|\mathbf{c}'_{[i]}\|_p \ \text{ s. t. } \ \mathbf{x}' = \mathbf{D}'\mathbf{c}'. \tag{4}$$

$\mathcal{P}'_{\ell_{0,p}}$ and $\mathcal{P}'_{\ell_{1,p}}$ can be converted to

$$\min_{\mathbf{c}'} \sum_i I(\|\mathbf{D}'\mathbf{c}'_{[i]}\|_p) \ \text{ s. t. } \ \mathbf{x}' = \mathbf{D}'\mathbf{c}', \tag{5}$$

and

$$\min_{\mathbf{c}'} \sum_i \|\mathbf{D}'\mathbf{c}'_{[i]}\|_p \ \text{ s. t. } \ \mathbf{x}' = \mathbf{D}'\mathbf{c}', \tag{6}$$

respectively.

Due to the structure of $\mathbf{D}'$ and the matrix norm we use here is element-wise norm, all the restricted isometry constants and block related coherence values defined in [2] of $\mathbf{D}'$ are identical to those of $\mathbf{D}$. Therefore, given a unique k-block sparse solution of $\mathcal{P}_{\ell_{0,p}}$ and $\mathcal{P}'_{\ell_{0,p}}$, the proof of the equivalence between $\mathcal{P}_{\ell_{1,p}}(\mathcal{P}'_{\ell_{1,p}})$ and $\mathcal{P}_{\ell_{0,p}}$ $(\mathcal{P}'_{\ell_{0,p}})$ under appropriate conditions can be can be constructed in a similar fashion as in [2].

1.2. Derivation of Reconstructed Block/Group Sparse Coding Algorithm

Continue from Section 2.3 in the main text. For clarity of presentation, we will again derive the reconstructed block/group sparse coding (R-BGSC) algorithm for only

*University of Florida, Gainesville, Florida, U. S. A.
†Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India

one single group of data. $\mathcal{P}'_{\ell_{1,p}}$ can be cast as a optimization problem that minimizes

$$\frac{1}{2}\|\mathbf{X}-\sum_{i\neq r}\mathbf{D}_{[i]}\mathbf{C}_{[i]}-\mathbf{D}_{[r]}\mathbf{C}_{[r]}\|_{\mathsf{F}}^2+\lambda\sum_i\left\|\mathbf{D}_{[i]}\mathbf{C}_{[i]}\right\|_p. \quad (7)$$

We now derive the crucial steps in optimization algorithm for $p = 2$ (as the $p = 1$ case is straightforward). Similar to the derivation in previous section, we first assume the norm $\left\|\mathbf{D}_{[r]}\mathbf{C}_{[r]}\right\|_{\mathsf{F}}$ is positive for optimal solution for $\mathbf{C}_{[r]}$. Taking the gradient of the objective function with respect to $\mathbf{C}_{[r]}$ and equating it to zero, we have

$$-\mathbf{D}_{[r]}^{\intercal}\mathbf{X}+\mathbf{D}_{[r]}^{\intercal}\sum_{i\neq r}\mathbf{D}_{[i]}\mathbf{C}_{[i]}+\mathbf{D}_{[r]}^{\intercal}\mathbf{D}_{[r]}\mathbf{C}_{[r]}+\lambda\mathbf{D}_{[r]}^{\intercal}\frac{\mathbf{D}_{[r]}\mathbf{C}_{[r]}}{\|\mathbf{D}_{[r]}\mathbf{C}_{[r]}\|_{\mathsf{F}}}=0. \quad (8)$$

Now denoting the first two terms with $-\mathbf{N}$ and computing the singular value decomposition of $\mathbf{D}_{[r]} = \mathbf{U}\mathbf{S}\mathbf{V}^{\intercal}$, we have

$$\mathbf{V}\mathbf{S}^2\mathbf{V}^{\intercal}\mathbf{C}_{[r]}+\lambda\mathbf{V}\mathbf{S}\frac{\mathbf{S}\mathbf{V}^{\intercal}\mathbf{C}_{[r]}}{\|\mathbf{S}\mathbf{V}^{\intercal}\mathbf{C}_{[r]}\|_{\mathsf{F}}}=\mathbf{N}. \quad (9)$$

Multiplying both sides of the above equation with $\mathbf{V}^{\intercal}$, and letting $\mathbf{Y} = \frac{\mathbf{S}\mathbf{V}^{\intercal}\mathbf{C}_{[r]}}{\|\mathbf{S}\mathbf{V}^{\intercal}\mathbf{C}_{[r]}\|_{\mathsf{F}}}$, $\kappa = \|\mathbf{S}\mathbf{V}^{\intercal}\mathbf{C}_{[r]}\|_{\mathsf{F}}$, and $\hat{\mathbf{N}} = \mathbf{V}^{\intercal}\mathbf{N}$, we have

$$(\kappa\mathbf{S}+\lambda\mathbf{S})\mathbf{Y}=\hat{\mathbf{N}} \quad \Longrightarrow \quad \mathbf{Y}=(\kappa\mathbf{S}+\lambda\mathbf{S})^{-1}\hat{\mathbf{N}},\ \text{ s. t. } \|\mathbf{Y}\|_{\mathsf{F}}=1. \quad (10)$$

Using the same method as in Section 2.2 in the main text, we can solve for κ first and then compute the iterate of $\mathbf{C}_{[r]}$.

1.3. Derivation of ICS-DL algorithm.

Continue from Eq. (17) in Section 2.4 in the main text, denoting the first two terms with $-\nu_i$, $\mathbf{c}_r\mathbf{c}_r^{\intercal}$ with t, and $\sum \mathbf{d}_j\mathbf{d}_j^{\intercal}$ with Φ_r, we have

$$t\mathbf{d}_r+\gamma\frac{\mathbf{d}_r}{\|\mathbf{d}_r\|_2}+\beta\Phi_r\mathbf{d}_r=\nu_i \implies t\mathbf{U}^{\intercal}\mathbf{d}_r+\gamma\frac{\mathbf{U}^{\intercal}\mathbf{d}_r}{\|\mathbf{U}^{\intercal}\mathbf{d}_r\|_2}+\beta\Sigma_{\Phi}\mathbf{U}^{\intercal}\mathbf{d}_r=\mathbf{U}^{\intercal}\nu_i, \quad (11)$$

where $\mathbf{U}\,\Sigma_{\Phi}\,\mathbf{U}^{\intercal}$ is the eigen decomposition of Φ_r and Σ_{Φ} is a diagonal matrix only contains the non-zero eigen-values of Φ_r and $\mathbf{U}$ are the corresponding eigen-vectors. Denoting $\frac{\mathbf{U}^{\intercal}\mathbf{d}_r}{\|\mathbf{U}^{\intercal}\mathbf{d}_r\|}$ by $\mathbf{y}$, $\|\mathbf{U}^{\intercal}\mathbf{d}_r\|_2$ by κ, and $\mathbf{U}^{\intercal}\nu_i$ by $\tilde{\nu}_i$, respectively, Eq. (11) then becomes

$$\kappa t\mathbf{y}+\gamma\mathbf{y}+\kappa\beta\,\Sigma_{\Phi}\mathbf{y}=\nu_i \implies \mathbf{y}=\left((\kappa t+\gamma)\mathbf{I}+\kappa\beta\Sigma_{\Phi}\right)^{-1}\tilde{\nu}_i,\ \text{ s. t. } \|\mathbf{y}\|_2=1. \quad (12)$$

We can use the same methods as in previous sections to solve for the iterate of $\mathbf{d}_r$, and if the solution κ is ≤ 0, we will set $\mathbf{d}_k = \mathbf{0}$.

Note that it is not uncommon to add a post-processing step to make atoms in $\mathbf{D}$ to have unit norms or simply requiring $\|\mathbf{d}_r\|_2$ to be 1. This changes the iterate of $\mathbf{d}_r$ to $\mathbf{d}_r = \left(t\mathbf{I}+\beta\Phi_r\right)^{-1}\nu_i$ as $\|\mathbf{d}_r\|_2 = 1$. Note that it is not uncommon to add a post-processing step to make atoms in $\mathbf{D}$ to have unit norms or simply constraint $\|\mathbf{d}_r\|_2$ to be 1. This makes the whole algorithm much more efficient as computing eigen-decomposition of a typically large matrix Φ_r can now be avoided.

2. Supplementary Experimental Results

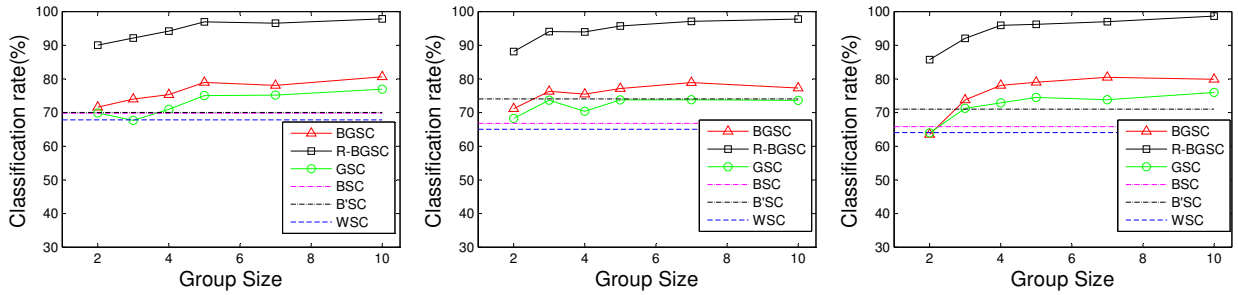
2.1. More results from face recognition

This section shows the supplementary results to Sec. 3.2 in the main text. Fig 1 shows the classification rates with dimensionality reduced to $m = \{100, \cdots, 500\}$. The "λ" for each algorithm are similar to what are listed in the main text because we normalized the vectors after the PCA projection.

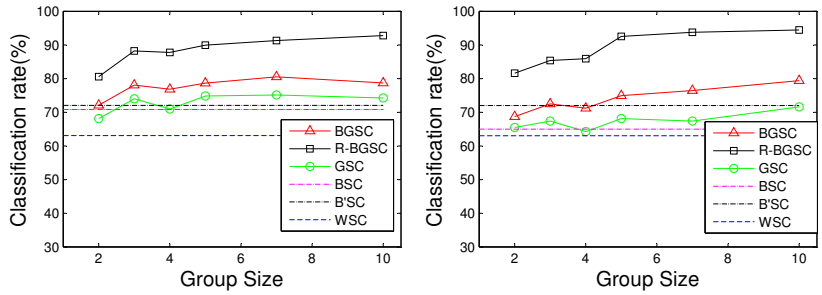
References

[1] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. *Advances in Neural Information Processing Systems*, 22:82–89, 2009.

[2] E. Elhamifar and R. Vidal. Block-sparse recovery via convex optimization. *Signal Processing, IEEE Transactions on*, PP(99):1, 2012.



(a) Data dimensionality m=500 (b) Data dimensionality m=400 (c) Data dimensionality m=300



(d) Data dimensionality m=200 (e) Data dimensionality m=100

Figure 1: Classification rates by different algorithms under different data dimensionality reduction ($m = 500 \cdots 100$).