Automatic Group Sparse Coding

Fei Wang[1], **Noah Lee**[1,2], **Jimeng Sun**[1], **Jianying Hu**[1] **and Shahram Ebadollahi**[1]
[1]Health Informatics Group, IBM T. J. Watson Research Center
[2]Department of Biomedical Engineering, Columbia University
–fwang,noahlee,jimeng,jyhu,ebad˝@us.ibm.com

Abstract

Sparse Coding (SC), which models the data vectors as sparse linear combinations over basis vectors (i.e., dictionary), has been widely applied in machine learning, signal processing and neuroscience. Recently, one specific SC technique, *Group Sparse Coding* (GSC), has been proposed to learn a common dictionary over multiple different groups of data, where the data groups are assumed to be pre-defined. In practice, this may not always be the case. In this paper, we propose *Automatic Group Sparse Coding* (AutoGSC), which can (1) discover the hidden data groups; (2) learn a common dictionary over different data groups; and (3) learn an individual dictionary for each data group. Finally, we conduct experiments on both synthetic and real world data sets to demonstrate the effectiveness of AutoGSC, and compare it with traditional sparse coding and Non-negative Matrix Factorization (NMF) methods.

Introduction

The linear decomposition of a signal (data vector) using a few atoms of a learned *dictionary*, or the *Sparse Coding* (SC) technique, has aroused considerable interests recently from various research fields such as audio processing (Févotte, Bertin, and Durrieu 2009), image denoising (Mairal, Elad, and Sapiro 2008), texture synthesis (Peyé 2009) and image classification (Bradley and Bagnell 2008). Different from traditional spectral decomposition methods such as *Principal Component Analysis* (PCA) and *Singular Value Decomposition* (SVD), SC (1) is usually additive, which results in a better representation ability; (2) does not require the learned bases to be orthogonal, which allows more flexibility to adapt the representation to the data set. In many real world applications (e.g., the ones we mentioned above), SC achieves state-of-the-art performance.

In traditional SC, each data vector is treated as an individual identity and the dictionary is learned over all these data vectors. Recently, (Bengio et al. 2009) pointed out that the SC procedure is just an intermediate step in creating a representation for a data group. For example, the data vectors could be the image descriptors or images, while one data group could be an image or image group. Clearly, the goal of SC is to learn how an image, not an image descriptor, is formed. Therefore (Bengio et al. 2009) proposed a novel technique called *Group Sparse Coding* (GSC), which can learn sparse representations at the group (image) level as well as a small overall dictionary (image descriptors).

One limitation of GSC is that it can only learn a common dictionary over all data groups. However, there should also be an individual dictionary associated with each data group, which makes those data groups different from each other. For example, in *electroencephalogram* (EEG) signal analysis when the data measured from several subjects under the same conditions (Lal et al. 2004; Lee and Choi 2009), each EEG signal contains some common as well as event (group) related frequency bands and regions. Moreover, in many cases, we only have the data vectors, while their associated group identities are hidden.

In this paper, we propose an *Automatic Group Sparse Coding* (AutoGSC) method, which assumes (1) there are hidden groups contained in the data set; (2) each data vector can be reconstructed using a sparse linear combination of both the common and group-specific dictionaries. We also proposed a Lloyd's style framework (Lloyd 1982) to learn both the data groups and those dictionaries. Specifically, it is worthwhile to emphasize the strength of AutoGSC.

- AutoGSC can learn hidden data groups automatically. In contrast, traditional GSC needs the data group identities to be pregiven.
- AutoGSC can learn an individual dictionary for each group, which contains group-specific discriminative information, while traditional GSC cannot.
- AutoGSC can also learn a common dictionary for all the groups as traditional GSC.

The rest of this paper is organized as follows. Section 2 introduces some notations and related works. The detailed algorithm and analysis is presented in section 3. Section 4 and 5 introduce the experimental results on synthetic and real world data, followed by the conclusions in section 6.

Background

Without the loss of generality, we assume the data instances are represented as vectors. Mathematically, we denote the observed data matrix as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$,

where $\mathbf{x}_i \in \mathbb{R}^d$ represents the i-th data instance vector. d is the data dimensionality, n is the number of data instances. Then the goal of sparse coding (Hoyer 2002; Mørup, Madsen, and Hansen 2008; Eggert and Korner 2004) is to obtain a sparse representation of the data vectors through a small set of basis vectors by minimizing

$$\mathcal{J}_0 = \left\|\mathbf{X} - \mathbf{F}\mathbf{G}^\top\right\|_F^2 + \lambda \sum_{i=1}^{n} \|\mathbf{G}_{i\cdot}\|_1 \tag{1}$$

where $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_k]$ is the dictionary matrix with $\mathbf{f}_i \in \mathbb{R}^d$ $(i = 1, 2, \cdots, k)$ being the i-th basis vector and k is the size of the dictionary. $\mathbf{G} \in \mathbb{R}^{n\times k}$ is the coding coefficient matrix. We use $G_{i\cdot}$ to denote the i-th row of $\mathbf{G}$, and

$$\|\mathbf{G}_{i\cdot}\|_1 = \sum_{j=1}^{k} |G_{ij}| \tag{2}$$

represents the ℓ_1 norm of $G_{i\cdot}$. $\lambda > 0$ is the tradeoff parameter. By expanding $\mathcal{J}_0$, we can obtain

$$\mathcal{J}_0 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} G_{ij}\mathbf{f}_j \right\|^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{k} |G_{ij}| \tag{3}$$

which shows that what SC actually seeks for is to approximate each data instance with a sparse linear combination of an appropriately learned dictionary. In this paper, we will concentrate on the *Non-Negative Sparse Coding* (NNSC) problem, i.e., $\mathbf{X} \geqslant 0, \mathbf{F} \geqslant 0, \mathbf{G} \geqslant 0$ (Here $\geqslant$ means elementwise nonnegativity). Then the optimization problem that NNSC tries to solve is

$$\min_{\mathbf{F}\geqslant 0,\ \mathbf{G}\geqslant 0} \mathcal{J}_0 \tag{4}$$

Unfortunately, problem (4) is not jointly convex with respect to both $\mathbf{F}$ and $\mathbf{G}$. However, it is convex with either of them with the other one fixed. Thus a common strategy for solving problem (4) is to adopt the *block coordinate descent* strategy (Bertsekas 1999), i.e., solve $\mathbf{F}$ and $\mathbf{G}$ alternatively with the other fixed until convergence.

When $\mathbf{G}$ is fixed, we can update $\mathbf{F}$ by *Multiplicative Updates* (Lee and Seung 2000) or *Projected Gradients* (Lin 2007). When $\mathbf{F}$ is fixed, the minimization of $\mathcal{J}_0$ with respect to $\mathbf{G}$ is an ℓ_1 regularized nonnegative least square problem. This type of problem can be solved by *LASSO* (Tibshirani 1996) or *Least Angle Regression* (LARS) (Efron et al. 2004; Eggert and Korner 2004).

However, as pointed out by (Eggert and Korner 2004), purely solving problem (4) may cause a scaling problem, as we can always scale up $\mathbf{F}$ and scale down $\mathbf{G}$ to get a lower cost function value. To overcome this problem, (Eggert and Korner 2004) proposed to minimize the following *normalization invariant* objective under nonnegativity constraints

$$\widehat{\mathcal{J}}_0 = \left\|\mathbf{X} - \widehat{\mathbf{F}}\mathbf{G}^\top\right\|_F^2 + \lambda \sum_{i=1}^{n} \|\mathbf{G}_{i\cdot}\|_1 \tag{5}$$

where $\widehat{\mathbf{F}} = [\mathbf{f}_1/\|\mathbf{f}_1\|, \mathbf{f}_2/\|\mathbf{f}_2\|, \cdots, \mathbf{f}_k/\|\mathbf{f}_k\|]$ is the normalized dictionary matrix, and $\|\mathbf{f}_i\| = \sqrt{\mathbf{f}_i^\top \mathbf{f}_i}$ is the Euclidean norm of $\mathbf{f}_i$. In this way, the objective function is evaluated on $\widehat{\mathbf{F}}$ and $\mathbf{G}$, and the scale of $\widehat{\mathbf{F}}$ is fixed.

Traditional SC treated each data instance as an individual and no data group information is considered. Sometimes it makes more sense to learn a group level sparse representation. Thus (Bengio et al. 2009) proposed *Group Sparse Coding* (GSC), which assumes that there are C hidden groups in $\mathbf{X}$. In the following, we use $\mathbf{X}_c = [\mathbf{x}_{c1}, \mathbf{x}_{c2}, \cdots, \mathbf{x}_{cn_c}] \in \mathbb{R}^{d\times n_c}$ to represent the c-th data group. $\mathbf{x}_{ci}$ is the i-th data instance of data group c. n_c is the size of the c-th group. Then the optimization problem GSC aims to solve is

$$\begin{aligned} &\min \sum_{c=1}^{C} \left[\left\|\mathbf{X}_c - \mathbf{F}\mathbf{G}_c^\top\right\|_F^2 + \lambda \sum_{i=1}^{n_c} \|\mathbf{G}_{ci\cdot}\|_p \right] + \gamma \sum_{j=1}^{k} \|\mathbf{F}_{\cdot j}\|_p \\ &s.t.\ \mathbf{F} \geqslant 0,\ \mathbf{G}_c \geqslant 0\ (c = 1, 2, \cdots, C) \end{aligned} \tag{6}$$

where $\mathbf{F} \in \mathbb{R}^{d\times k}$ is a common shared dictionary over all groups, $\mathbf{G}_c \in \mathbb{R}^{n_c\times k}$ is the coding coefficient matrix for group c. $\mathbf{G}_{ci\cdot}$ is the i-th row of $\mathbf{G}_c$, $\mathbf{F}_{\cdot j}$ is the j-th column of $\mathbf{F}$. $\|\mathbf{a}\|_p$ is the general ℓ_p norm of a vector $\mathbf{a}$, and in this paper, we will consider $p = 1$ because it is the most popular choice in sparse coding. Here the regularization term of $\mathbf{F}$ plays a similar role as the normalization of $\mathbf{F}$ on Eq. (5).

By solving problem (6), GSC learns a sparse representation on group level as well as a shared dictionary. However, GSC assumes the data group identities are pre-given and it can only learn a common dictionary. However, in many real world applications, (1) the data group identities are hidden and (2) we want to know the group-specific dictionaries, as these individual dictionaries can help us to capture the discrimination information contained in different data groups.

Based on the above considerations, in this paper, we propose *Automatic Group Sparse Coding* (AutoGSC), which can (1) discover the hidden data groups; (2) learn a common dictionary over different data groups; (3) learn an individual dictionary for each data group. The algorithm details will be introduced in the next section.

Automatic Group Sparse Coding

In AutoGSC, we also assume there are C groups contained in the data set with the c-th group $\mathbf{X}_c$. Then we assume there is a shared dictionary $\mathbf{F}^S \in \mathbb{R}^{d\times k^S}$ over all C groups, where k^S is the dictionary size. Also there is an individual dictionary $\mathbf{F}_c^I \in \mathbb{R}^{d\times k_c^I}$ for each group c, with the dictionary size k_c^I. Then the problem that AutoGSC tries to solve is

$$\begin{aligned} &\min \sum_{c} \left\|\mathbf{X}_c - \mathbf{F}^S\mathbf{G}_c^{S\top} - \mathbf{F}_c^I\mathbf{G}_c^{I\top}\right\|_F^2 + \sum_{c} \left[\gamma_I \phi\left(\mathbf{G}_c^I\right) + \gamma_S \phi\left(\mathbf{G}_c^S\right)\right] \\ &s.t.\ \mathbf{F}^S \geqslant 0,\ \forall c = 1, 2, \cdots, C,\ \mathbf{F}_c^I \geqslant 0, \mathbf{G}_c^I \geqslant 0, \mathbf{G}_c^S \geqslant 0 \end{aligned} \tag{7}$$

where the variables we want to solve in the above problem include $\mathbf{F}^S, \{\mathbf{F}_c^I\}_{c=1}^C, \{\mathbf{G}_c^S\}_{c=1}^C, \{\mathbf{G}_c^I\}_{c=1}^C$, as well as the data group identities. The first term of the objective measures the total reconstruction error (using matrix Frobenius norm) of the data set from those common and individual dictionaries. $\mathbf{G}_c^S \in \mathbb{R}^{n_c\times k^S}$ is the reconstruction coefficient matrix on the group-shared dictionary $\mathbf{F}^S$. $\mathbf{G}_c^I$ is the reconstruction coefficient matrix on c-th group-specific dictionary

$\mathbf{G}_c^I \in \mathbb{R}^{n_c \times k_c^I}$. The second term imposes some regularizations on the coding coefficients. As what we care is sparse coding here, we make the following specific assumptions

$$\phi\left(\mathbf{G}_c^I\right) = \sum_{i=1}^{n_c} \left\|\mathbf{G}_{ci\cdot}^I\right\|_1 \tag{8}$$

$$\phi\left(\mathbf{G}_c^S\right) = \sum_{i=1}^{n_c} \left\|\mathbf{G}_{ci\cdot}^S\right\|_1 \tag{9}$$

where $\mathbf{G}_{ci\cdot}^S$ and $\mathbf{G}_{ci\cdot}^I$ is the i-th row of $\mathbf{G}_c^S$ and $\mathbf{G}_c^I$. $\|\cdot\|_1$ represents the vector ℓ_1 norm defined as in Eq.(2). Similar as in Eq.(5), we solve the following dictionary normalization invariant version of problem (7) instead

$$\min \sum_c \left\|\mathbf{X}_c - \widehat{\mathbf{F}}^S \mathbf{G}_c^{S\top} - \widehat{\mathbf{F}}_c^I \mathbf{G}_c^{I\top}\right\|_F^2 + \sum_c \left[\gamma_I \phi\left(\mathbf{G}_c^I\right) + \gamma_S \phi\left(\mathbf{G}_c^S\right)\right]$$
$$s.t.\ \mathbf{F}^S \geqslant 0,\ \forall c = 1, 2, \cdots, C,\ \mathbf{F}_c^I \geqslant 0, \mathbf{G}_c^I \geqslant 0, \mathbf{G}_c^S \geqslant 0 \tag{10}$$

where $\widehat{\mathbf{F}}^S = [\mathbf{f}_1^S/\|\mathbf{f}_1^S\|, \mathbf{f}_2^S/\|\mathbf{f}_2^S\|, \cdots, \mathbf{f}_{n^S}^S/\|\mathbf{f}_{n^S}^S\|]$, $\mathbf{f}_i^S$ is the i-th column of $\mathbf{F}^S$. $\widehat{\mathbf{F}}_c^I = [\mathbf{f}_{c1}^I/\|\mathbf{f}_{c1}^I\|, \mathbf{f}_{c2}^I/\|\mathbf{f}_{c2}^I\|, \cdots, \mathbf{f}_{ck_c^I}^I/\|\mathbf{f}_{ck_c^I}^I\|]$, $\mathbf{f}_{ci}^I$ is the i-th column of $\mathbf{F}_c^I$. As we mentioned, we need to solve both the data group identities as well as $\widehat{\mathbf{F}}^S, \{\widehat{\mathbf{F}}_c^I\}_{c=1}^C, \{\mathbf{G}_c^S\}_{c=1}^C, \{\mathbf{G}_c^I\}_{c=1}^C$, which is not an easy task. However, if we define

$$\widehat{\mathbf{F}}_c = [\widehat{\mathbf{F}}^S, \widehat{\mathbf{F}}_c^I] \tag{11}$$

$$\mathbf{G}_c = [\mathbf{G}_c^S, \mathbf{G}_c^I] \tag{12}$$

and assume $\gamma_I = \gamma_S = \gamma$, then we can rewrite the objective of problem (7) as

$$\sum_c \left[\left\|\mathbf{X}_c - \widehat{\mathbf{F}}_c \mathbf{G}_c^\top\right\|_F^2 + \gamma\phi\left(\mathbf{G}_c\right)\right] \tag{13}$$
$$= \sum_c \sum_{\mathbf{x}_i \in \pi_c} \left[\left\|\mathbf{x}_i - \sum_j G_{cij} \widehat{\mathbf{F}}_{c\cdot j}\right\|^2 + \gamma \sum_j |G_{cij}|\right]$$

where π_c is the c-th data group, $\widehat{\mathbf{F}}_{c\cdot j}$ is the j-th column of $\widehat{\mathbf{F}}_c$, G_{cij} is the (i,j)-th entry of $\mathbf{G}_c$. This is very similar to the problem of *Vector Quantization* (VQ) (Lloyd 1982). The main difference is that in AutoGSC, we use C dictionaries $\{\widehat{\mathbf{F}}_c\}_{c=1}^C$ to quantize those data vectors, instead of using C vectors as in traditional VQ. Based on this observation, we propose a Lloyd style algorithm (Lloyd 1982) to solve the problem, which alternates between the following two steps:

- Solving problem (10) to get the dictionaries as well as the coding coefficients with given data group identities.
- Estimating data group identities using the current dictionaries and codes.

In the following we will introduce how these two steps proceed in detail.

Obtaining the Dictionaries

Given the data group identities, we can solve problem (10) to get the dictionaries as well as the coding coefficients. More formally, there are four groups of variables in problem (10): $\widehat{\mathbf{F}}^S, \{\mathbf{G}_c^S\}_{c=1}^C, \{\widehat{\mathbf{F}}_c^I\}_{c=1}^C, \{\mathbf{G}_c^I\}_{c=1}^C$. We adopt an alternating scheme to update them. Similar to traditional SC techniques, if we fix the others, the updating of $\mathbf{G}_c^S$ or $\mathbf{G}_c^I$ ($\forall\, c = 1, 2, \cdots, C$) would just involve an ℓ_1 regularized nonnegative least square regression problem, which can be solved using LASSO (Tibshirani 1996) or LARS (Efron et al. 2004; Mørup, Madsen, and Hansen 2008). For $\mathbf{F}^S$, we can update it using the following update rule

$$\mathbf{F}^S \longleftarrow \mathbf{F}^S \odot \frac{\sum_{c=1}^C \left[\mathbf{A}_c^S + \widehat{\mathbf{F}}^S \mathrm{diag}\left(\mathbf{1}^\top \left(\mathbf{B}_c^S \odot \widehat{\mathbf{F}}^S\right)\right)\right]}{\sum_{c=1}^C \left[\mathbf{B}_c^S + \widehat{\mathbf{F}}^S \mathrm{diag}\left(\mathbf{1}^\top \left(\mathbf{A}_c^S \odot \widehat{\mathbf{F}}^S\right)\right)\right]} \tag{14}$$

where

$$\mathbf{A}_c^S = \mathbf{X}_c \mathbf{G}_c^S \tag{15}$$

$$\mathbf{B}_c^S = \widehat{\mathbf{F}}^S \mathbf{G}_c^{S\top} \mathbf{G}_c^S + \widehat{\mathbf{F}}_c^I \mathbf{G}_c^{I\top} \mathbf{G}_c^S \tag{16}$$

$\odot$ represents the matrix elementwise product, and $-$ means matrix elementwise division. For $\mathbf{F}_c^I$ ($c = 1, 2, \cdots, C$), we can update it with

$$\mathbf{F}_c^I \longleftarrow \mathbf{F}_c^I \odot \frac{\mathbf{A}_c^I + \widehat{\mathbf{F}}^S \mathrm{diag}\left(\mathbf{1}^\top \left(\mathbf{B}_c^I \odot \widehat{\mathbf{F}}^S\right)\right)}{\mathbf{B}_c^I + \widehat{\mathbf{F}}^S \mathrm{diag}\left(\mathbf{1}^\top \left(\mathbf{A}_c^I \odot \widehat{\mathbf{F}}^S\right)\right)} \tag{17}$$

where

$$\mathbf{A}_c^I = \mathbf{X}_c \mathbf{G}_c^I \tag{18}$$

$$\mathbf{B}_c^I = \widehat{\mathbf{F}}^S \mathbf{G}_c^{S\top} \mathbf{G}_c^I + \widehat{\mathbf{F}}_c^I \mathbf{G}_c^{I\top} \mathbf{G}_c^I \tag{19}$$

The correctness of the updating rules Eq.(14) and Eq.(17) are guaranteed by the following theorem.

Theorem. *If the update rule of $\widehat{\mathbf{F}}^S$ and $\{\widehat{\mathbf{F}}_c^I\}_{c=1}^C$ in Eq.(14) and Eq.(17) converges, then the final solution satisfies the Karush-Kuhn-Tucker (KKT) optimality condition.*

Proof. See Appendix.

Obtaining the Group Identities

As we can see from Eq.(13), what AutoGSC actually does is to *quantize* the data space using C dictionaries $\widehat{\mathbf{F}}_1, \widehat{\mathbf{F}}_2, \cdots, \widehat{\mathbf{F}}_C$. The error for quantizing $\mathbf{x}_i$ with dictionary $\widehat{\mathbf{F}}_c$ can be measured by

$$\mathcal{Q}(\mathbf{x}_i, \widehat{\mathbf{F}}_c) = \min_{\mathbf{g}_{c_i}} \|\mathbf{x}_i - \widehat{\mathbf{F}}_c \mathbf{g}_{c_i}\|^2 + \gamma |\mathbf{g}_{c_i}|_1 \tag{20}$$

and the group identity of $\mathbf{x}_i$ can be predicted as

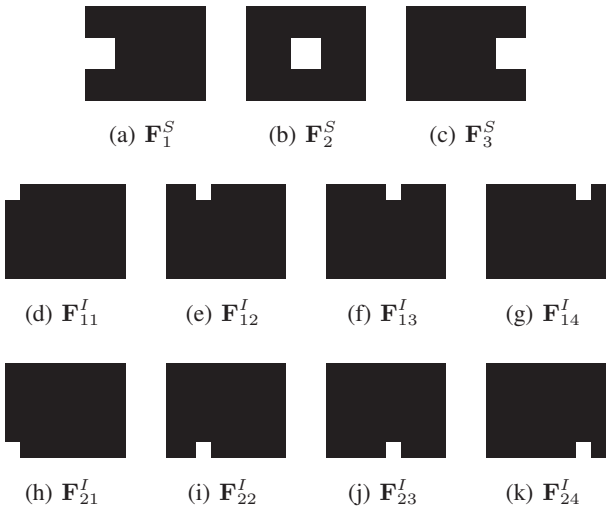
$$GI(\mathbf{x}_i) = arg\min_c \mathcal{Q}(\mathbf{x}_i, \widehat{\mathbf{F}}_c) \tag{21}$$

A Synthetic Example

In this section we will introduce a set of experiments to validate the effectiveness of the proposed AutoGSC algorithm.

First we shall show a synthetic example. The data set we use here is a set of images of size 30×40 constituting two groups. Both groups have three common basis images shown in Fig.1(a)(b)(c), where we use dark colors to represent value zero, and bright colors to represent value 1. Group 1 has four individual basis images shown in Fig.1(d)-(g).

Group 2 has four individual basis images shown in Fig.1(h)-(k). The set of images used in our experiments are generated by a random combination of a pair of common and individual basis images plus some uniform random noise with values in $[0, 0.1]$. Fig.2 illustrates examples of the training images, where the top row belongs to the first group, while bottom row belongs to the second group.



(a) $\mathbf{F}_1^S$ (b) $\mathbf{F}_2^S$ (c) $\mathbf{F}_3^S$

(d) $\mathbf{F}_{11}^I$ (e) $\mathbf{F}_{12}^I$ (f) $\mathbf{F}_{13}^I$ (g) $\mathbf{F}_{14}^I$

(h) $\mathbf{F}_{21}^I$ (i) $\mathbf{F}_{22}^I$ (j) $\mathbf{F}_{23}^I$ (k) $\mathbf{F}_{24}^I$

Figure 1: Common and individual dictionaries.

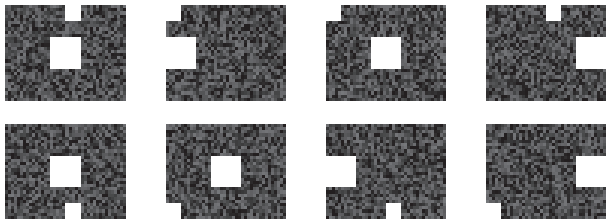
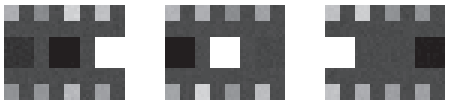


Figure 2: Examples of the training data.

Fig.3 illustrates the three basis learned using simple Nonnegative Matrix Factorization (NMF) (Lin 2007), which obtains basis images $\mathbf{F}$ by minimizing $\|\mathbf{X} - \mathbf{F}\mathbf{G}^\top\|_F^2$ using projected gradient with randomly initialized $\mathbf{F}$. From the figure we can see that the three common basis contained in the data set are correctly learned, however, they are mixed with the individual basis as traditional NMF does not have the scheme to discriminate common and individual basis. The similar phenomenon can be observed when we apply simple nonnegative sparse coding (Eggert and Korner 2004), which minimizes $\|\mathbf{X}-\mathbf{F}\mathbf{G}^\top\|_F^2+\lambda\sum_{ij}|G_{ij}|$ with normalization invariant updates.

Fig.5 illustrate the results of running unsupervised Group-NMF (Lee and Choi 2009) on the data set, i.e., solve problem (7) with Lloyd's framework without sparsity regularization and basis wise normalization. These figures demonstrate that both the common and individual patterns are mixed up in this case.



(a) $\mathbf{F}_1$ (b) $\mathbf{F}_2$ (c) $\mathbf{F}_3$

Figure 3: Dictionary learned by Nonnegative Matrix Factorization (Lin 2007).

(a) $\mathbf{F}_1$ (b) $\mathbf{F}_2$ (c) $\mathbf{F}_3$

Figure 4: Dictionary learned by group NMF (Lee and Choi 2009).

Fig.6 shows the learned basis images AutoGSC proposed in this paper. We can see that both the common individual basis images, are correctly learned. Note that we use the same (random) initializations for $\mathbf{F}^S, \mathbf{F}^I, \mathbf{G}^S, \mathbf{G}^I$ as well as the data group identities to obtain the results shown in Fig.5 and Fig.6. The number of groups is set to 2.

Fig.7 demonstrates the convergence curve of running AutoGSC on our synthetic data set, which shows that our algorithm can converge within about 30 steps in this case.

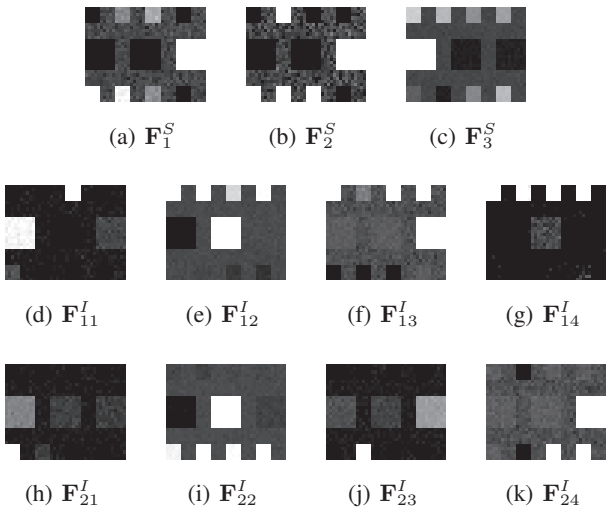
A Case Study

In this section we will apply our AutoGSC algorithm to a real world scenario of medical informatics.

Specifically, effective patient utilization management is an important issue for medical care delivery. Here we refer to *utilization* as different types of patient visits, such as visits to a Primary Care Physician (PCP), specialist, independent lab, in/out-patient hospital, etc. Usually management on high-utilization patients receives more attention as these patients consume more resources. A well accepted fact in medical informatics is that *20% of the patients incur 80% of the cost*. In the following, we will make use of AutoGSC to investigate the clinical characteristics of the high utilization patient population, i.e., detect the disease groups as well as the common and individual representative diseases.

The data set we use consists of patients' clinical records, including clinical characteristics, demographic features, utilizations, medication history, for a pool of over 131k patients over one year period. We compute the total number of visits for each patient, and use that count as the indication of the patient utilization level. We plot the histogram of patient visits in Fig.8, and with medical expert assistance, we select the cutoff point to be 100, i.e., a patient is considered to incur high utilization if the number of his yearly visits is larger than 100. In this way, we obtain a pool of 216 patients. Then we use the *HCC codes*[1] to represent the patient

[1]HCC stands for *Hierarchical Condition Category* (Pope et al. 2000), which can be viewed as a grouping of ICD9 (*International*

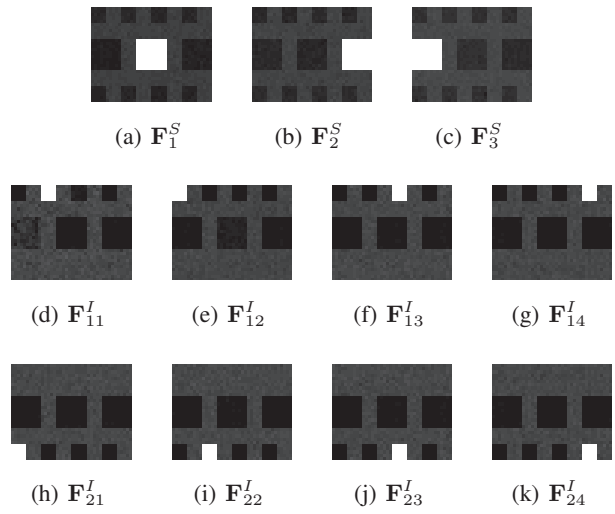


(a) $\mathbf{F}_1^S$ (b) $\mathbf{F}_2^S$ (c) $\mathbf{F}_3^S$

(d) $\mathbf{F}_{11}^I$ (e) $\mathbf{F}_{12}^I$ (f) $\mathbf{F}_{13}^I$ (g) $\mathbf{F}_{14}^I$

(h) $\mathbf{F}_{21}^I$ (i) $\mathbf{F}_{22}^I$ (j) $\mathbf{F}_{23}^I$ (k) $\mathbf{F}_{24}^I$

Figure 5: Common and individual dictionaries learned by GroupNMF (Lee and Choi 2009).



(a) $\mathbf{F}_1^S$ (b) $\mathbf{F}_2^S$ (c) $\mathbf{F}_3^S$

(d) $\mathbf{F}_{11}^I$ (e) $\mathbf{F}_{12}^I$ (f) $\mathbf{F}_{13}^I$ (g) $\mathbf{F}_{14}^I$

(h) $\mathbf{F}_{21}^I$ (i) $\mathbf{F}_{22}^I$ (j) $\mathbf{F}_{23}^I$ (k) $\mathbf{F}_{24}^I$

Figure 6: Common and individual dictionaries learned by AutoGSC.

diagnosis features. In this way, we obtain a 195×216 patient matrix $\mathbf{X}$ (as we have 195 distinct HCC codes), with

$$X_{ij} = \begin{cases} 1, & \text{if patient } j \text{ was diagnosed with HCC code } i \\ 0, & \text{otherwise} \end{cases}$$

We run AutoGSC on $\mathbf{X}$ with random initializations, and set the number of groups to be 3. We set the number of common as well as individual condition basis to be 5 (here each condition basis is a 195 dimensional vector). The learned basis are very sparse, i.e., most of the elements on the learned condition basis vectors are zero. Here we refer to the conditions with nonzero values in the basis vectors as *active conditions*. Table 1 illustrates the common active conditions.

Table 1: Common Active conditions

HCC code	Description
HCC166	Major Symptoms, Abnormalities
HCC179	Post-Surgical States/Aftercare/Elective
HCC167	Minor Symptoms, Signs, Findings
HCC183	Screening/Observation/Special Exams
HCC162	Other Injuries

Table 2 shows the active conditions found in group 1, which are different types of cancers. Table 3 illustrates the active conditions found in group 2, which are mainly heart conditions. Table 4 shows the active conditions found in group 3, which are related to some major surgeries such as organ transplant. From these results we can clearly observe the major condition groups that may lead to high utilization. Such insights will be very useful in optimizing care delivery to patients and reducing the associated cost.

Statistical Classification of conditions and Related Health Problems, 9th ed.) diagnosis codes for better healthcare management.

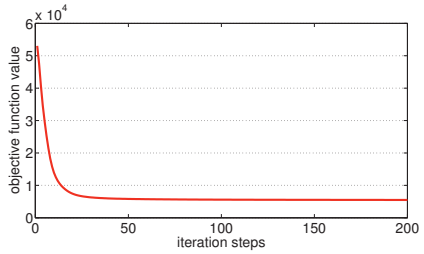


Figure 7: The objective function value vs. number of iterations plot on the synthetic data set.

Conclusion

In this paper we proposed *Automatic Group Sparse Coding* (*AutoGSC*). Different from traditional group sparse coding, AutoGSC can (1) learn both common as well as individual basis for all data groups; (2) automatically find the hidden data groups. We provide experimental results on applying AutoGSC to a synthetic data set. Finally we also use it to discover representative condition groups for high utilization patient population, which demonstrates the effectiveness of AutoGSC in real heatlcare applications.

Appendix

Following the standard theory of constrained optimization, we introduce the Lagrangian multipliers $\boldsymbol{\alpha} = [\alpha_{ij}] \in$

Table 2: Active conditions in Group 1

HCC code	Description
HCC312	Breast, Prostate, and Other Cancers and Tumors
HCC311	Colorectal, Bladder, and Other Cancers
HCC310	Lymphoma and Other Cancers
HCC309	Lung and Other Severe Cancers

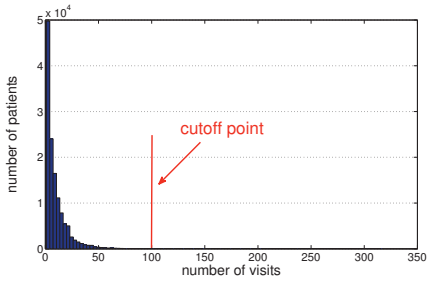


Figure 8: The histogram of the patient visits.

Table 3: Active conditions in Group 2

HCC code	Description
HCC080	Congestive Heart Failure
HCC079	Cardio-Respiratory Failure and Shock
HCC092	Specified Heart Arrhythmias
HCC091	Hypertension
HCC092	Specified Heart Arrhythmias

$\mathbb{R}^{d\times k^S}$ and $\boldsymbol{\beta}_c \in \mathbb{R}^{d\times n_c^I}$ $(c = 1, 2, \cdots, C)$ and construct the following Lagrangian function

$$\mathcal{L}=\sum_{c=1}^{C}\left[\left\|\mathbf{X}_c-\widehat{\mathbf{F}}^S\mathbf{G}_c^{S\top}-\widehat{\mathbf{F}}_c^I\mathbf{G}_c^{I\top}\right\|_F^2-tr\left(\widehat{\mathbf{F}}_c^I\boldsymbol{\beta}_c^\top\right)\right]-tr\left(\widehat{\mathbf{F}}^S\boldsymbol{\alpha}^\top\right)$$

Taking the first order derivative, we have

$$\left.\frac{\partial\mathcal{L}}{\partial\mathbf{F}^S}\right|_{uv}=-\frac{2}{\|\mathbf{F}_{\cdot v}^S\|^2}\sum_{c=1}^{C}\left[\left(\mathbf{A}_c^S-\mathbf{B}_c^S\right)_{uv}\|\mathbf{F}_{\cdot v}^S\| - \left(\widehat{\mathbf{F}}^S\text{diag}\left(\mathbf{1}^\top\left[\left(\mathbf{A}_c^S-\mathbf{B}_c^S\right)\odot\mathbf{F}^S\right]\right)\right)\right]_{uv}-\alpha_{uv} \quad (22)$$

where $\mathbf{A}_c^S$ and $\mathbf{B}_c^S$ are defined as in Eq.(15) and Eq.(16)Fixing the other variables and setting $\partial\mathcal{L}/\partial\mathbf{F}^S = 0$, we have

$$\alpha_{uv}=-\frac{2}{\|\mathbf{F}_{\cdot v}^S\|^2}\sum_{c=1}^{C}\left[\left(\mathbf{A}_c^S-\mathbf{B}_c^S\right)_{uv}\|\mathbf{F}_{\cdot v}^S\| - \left(\widehat{\mathbf{F}}^S\text{diag}\left(\mathbf{1}^\top\left[\left(\mathbf{A}_c^S-\mathbf{B}_c^S\right)\odot\mathbf{F}^S\right]\right)\right)\right]_{uv} \quad (23)$$

the KKT complementary condition for the nonnegativity of $\mathbf{F}$ is

$$\alpha_{uv}\mathbf{F}_{uv}=-\frac{2}{\|\mathbf{F}_{\cdot v}^S\|^2}\sum_{c=1}^{C}\left[\left(\mathbf{A}_c^S-\mathbf{B}_c^S\right)_{uv}\|\mathbf{F}_{\cdot v}^S\| - \left(\widehat{\mathbf{F}}^S\text{diag}\left(\mathbf{1}^\top\left[\left(\mathbf{A}_c^S-\mathbf{B}_c^S\right)\odot\mathbf{F}^S\right]\right)\right)\right]_{uv}\mathbf{F}_{uv}=0 \quad (24)$$

Table 4: Active conditions in Group 3

HCC code	Description
HCC174	Major Organ Transplant Status
HCC179	Post-Surgical States/Aftercare/Elective
HCC160	Internal Injuries
HCC023	Disorders of Fluid/Electrolyte/Acid-Base Balance
HCC044	Severe Hematological Disorders

which is equivalent to

$$\mathbf{F}^S=\mathbf{F}^S\odot\frac{\sum_{c=1}^{C}\left[\mathbf{A}_c^S+\widehat{\mathbf{F}}^S\text{diag}\left(\mathbf{1}^\top\left(\mathbf{B}_c^S\odot\widehat{\mathbf{F}}^S\right)\right)\right]}{\sum_{c=1}^{C}\left[\mathbf{B}_c^S+\widehat{\mathbf{F}}^S\text{diag}\left(\mathbf{1}^\top\left(\mathbf{A}_c^S\odot\widehat{\mathbf{F}}^S\right)\right)\right]} \quad (25)$$

This is exactly the same as in Eq.(14) when the iteration converges. Similarly, we have

$$\left.\frac{\partial\mathcal{J}_1}{\partial\mathbf{F}_c^I}\right|_{uv}=\left(\mathbf{A}_c^I-\mathbf{B}_c^I\right)_{uv}\frac{\|\mathbf{F}_{c\cdot v}^I\|}{\|\mathbf{F}_{c\cdot v}^I\|^2} - \left[\widehat{\mathbf{F}}_c^I\text{diag}\left(\mathbf{1}^\top\left[\left(\mathbf{A}_c^I-\mathbf{B}_c^I\right)\odot\mathbf{F}_c^I\right]\right)\right]_{uv}\frac{1}{\|\mathbf{F}_{\cdot v}^S\|^2}-\beta_{cuv}$$

and it can be easily validated that Eq.(17) satisfies the KKT complementary condition when converges. □

References

Bengio, S.; Pereira, F.; Singer, Y.; and Strelow, D. 2009. Group sparse coding. In *NIPS 22*.

Bertsekas, D. P. 1999. *Nonlinear Programming*. Athena Scientific, 2nd edition.

Bradley, D. M., and Bagnell, J. A. 2008. Differentiable sparse coding. In *NIPS 21*, 113–120.

Efron, B.; Hastie, T.; Johnstone, L.; and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics* 32:407–499.

Eggert, J., and Korner, E. 2004. Sparse coding and nmf. In *Proceedings of IEEE International Joint Conference on Neural Networks*, volume 4, 2529–2533.

Févotte, C.; Bertin, N.; and Durrieu, J. L. 2009. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation* 21(3):793–830.

Hoyer, P. O. 2002. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, 557–565.

Lal, T. N.; Schröder, M.; Hinterberger, T.; Weston, J.; Bogdan, M.; Birbaumer, N.; and Schölkopf, B. 2004. Support vector channel selection in BCI. *IEEE Trans Biomed Eng* 51(6):1003–1010.

Lee, H., and Choi, S. 2009. Group nonnegative matrix factorization for eeg classification. In *AISTATS 12*, 320–327.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 556–562.

Lin, C.-J. 2007. Projected gradient methods for non-negative matrix factorization. *Neural Computation* 2756–2779.

Lloyd, S. P. 1982. Least squares quantization in pcm. *IEEE Trans. on Information Theory* 28(2):129–137.

Mairal, J.; Elad, M.; and Sapiro, G. 2008. Sparse representation for color image restoration. *IEEE Trans. on Image Processing* 17(1):53–69.

Mørup, M.; Madsen, K. H.; and Hansen, L. K. 2008. Approximate l0 constrained non-negative matrix and tensor factorization. In *Proceedings of Int'l Symp. on Circuits and Systems*, 1328–1331.

Peyé, G. 2009. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision* 34(1):17–31.

Pope, G.; Ellis, R. P.; Ash, A. S.; Ayanian, J. Z.; Bates, D. W.; Burstin, H.; Iezzoni, L. I.; Marcantonio, E.; and Wu, B. 2000. Details for diagnostic cost group hierarchical condition category models for medicare risk adjustment. *Research Report*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. of the Royal Statistical Society (Series B)* 58:267–288.