# Non-Negative Semi-Supervised Learning

**Changhu Wang**[*]
MOE-MS Key Lab of MCC
Univ. of Sci. and Tech. of China
Heifei 230027, China

**Shuicheng Yan**
Dept. of Elec. and Comp. Eng.
National University of Singapore
117576, Singapore

**Lei Zhang**[1], **Hong-Jiang Zhang**[2]
[1]Microsoft Research Asia
[2]Microsoft Adv. Tech. Center
Beijing 100190, China

## Abstract

The contributions of this paper are three-fold. First, we present a general formulation for reaping the benefits from both non-negative data factorization and semi-supervised learning, and the solution naturally possesses the characteristics of sparsity, robustness to partial occlusions, and greater discriminating power via extra unlabeled data. Then, an efficient multiplicative updating procedure is proposed along with its theoretic justification of the algorithmic convergency. Finally, the tensorization of this general formulation for non-negative semi-supervised learning is also briefed for handling tensor data of arbitrary order. Extensive experiments compared with the state-of-the-art algorithms for non-negative data factorization and semi-supervised learning demonstrate the algorithmic properties in sparsity, classification power, and robustness to image occlusions.

## 1 INTRODUCTION

Motivated by the psychological and physiological evidence for parts-based representations in the brain (Lee & Seung, 1999), recently techniques for non-negative and sparse representation have been well studied for finding non-negative bases with few nonzero elements. Non-negative matrix factorization (NMF) (Lee & Seung, 1999), as a pioneering work for such a purpose, has shown its powerful capability for parts-based representations of images and other types of data. NMF is distinguished from other holistic-based methods by its use of non-negativity constraints, which ensure that an image could only be formed from non-negative bases in a non-subtractive way, and therefore lead to a parts-based representation.

Following the work of NMF, many algorithms have been proposed for non-negative data decomposition and classification. Li et al. (2001) imposed extra constraints to reinforce the basis sparsity of NMF; also matrix-based NMF has been extended to non-negative tensor factorization (NTF) (Hazan et al., 2005; Shashua & Hazan, 2005) for handling the data encoded as high-order tensors. Wang et al. proposed the Fisher-NMF (2004), which was further studied by Kotsia et al. (2007), by adding an extra term of scatter difference to the objective function of NMF. Tao et al. (2005) proposed to employ local rectangle binary features for image reconstruction. Recently, Yang et al. (2008a) proposed a general solution for supervised non-negative graph embedding by integrating the characteristics of both intrinsic and penalty graphs (Yan et al., 2007) with non-negative data factorization.

Most of these algorithms proposed for non-negative data factorization are unsupervised. Among the supervised ones, the supervised non-negative graph embedding proposed in (Yang et al., 2008a), although with the superiority over Fisher-NMF, suffers from the high computational cost caused by calculating the inverse of the so-called $M$-matrix (Yang et al., 2008a), which greatly limits its practical applications. Beyond supervised learning, many recent research (Zhou et al., 2003; Zhu et al., 2003; Belkin et al., 2004; Belkin et al., 2006; Cai et al., 2007) shows that the learning process may greatly benefit from the unlabeled data, which are often relatively easy to obtain in practice. A detailed literature survey on semi-supervised learning is referred to (Zhu, 2005). A natural question to ask is whether we can design an algorithm with three characteristics: 1) the derived solution is non-negative and sparse, and hence robust to partial image occlusions; 2) the formulation may well

[*]Changhu Wang performed this work while being a Research Engineer at the Department of Electrical and Computer Engineering, National University of Singapore. This work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research Grant NRF2008IDM-IDM004-029.

Appearing in Proceedings of the $12^{th}$ International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

utilize the unlabeled data for achieving greater discriminating power; and 3) the procedure to obtain such a solution is efficient, ideally again based on the elegant multiplicative updating rule.

This work is dedicated to designing such a data factorization algorithm with the above-mentioned three characteristics, referred to as non-negative semi-supervised learning (N$^2$S$^2$L). First, we present a general formulation for reaping the benefits from both non-negative data factorization (sparsity and robustness to partial occlusion) and semi-supervised learning (greater discriminating power via extra unlabeled data). Then, an efficient multiplicative updating procedure is proposed along with its theoretic justification of the algorithmic convergency. Finally, the tensorization of this general formulation for non-negative semi-supervised learning is also briefed for handling tensor data of arbitrary order.

The remainder of this paper is organized as follows. In Section 2, we introduce the details of N$^2$S$^2$L algorithm. In Section 3, the N$^2$S$^2$L algorithm is further generalized for handling tensor data of arbitrary order. The comparison experiments are demonstrated in Section 4.

## 2 NON-NEGATIVE SEMI-SUPERVISED MATRIX FACTORIZATION

In this section, we introduce the math formulation and its iterative multiplicative updating rule for the non-negative semi-supervised matrix factorization problem, where each datum is represented by a vector. We assume that the training data are given as $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in \mathbb{R}^m$, and $N$ is the total number of training samples. Portion of the data are labeled as $c_i \in \{1, ..., N_c\}$, where $N_c$ is the class number. Denote the sample number of the $c$th class as $n_c$. Note that we utilize in this work the following rule to facilitate presentation: for any matrix $A$, its corresponding lowercase version $a_i$ means the $i$th column vector of $A$, and $A_{ij}$ denotes the element of $A$ at the $i$th row and $j$th column.

### 2.1 PROBLEM FORMULATION

To achieve the ultimate target of N$^2$S$^2$L, the objective function need involve different components: 1) the component to guide the parts-based data decomposition; 2) the component to guarantee the separability of the labeled data; and 3) the component on the extra regularization from both labeled and unlabeled data.

#### 2.1.1 Objective for Non-negative Data Reconstruction

Non-negative matrix factorization (NMF) algorithm uses two non-negative matrices, *i.e.*, one lower-rank basis matrix and one coefficient matrix, to reconstruct the original data matrix. Its objective function is,

$$\min_{U,V} \|X - UV^T\|_F^2, \;\; s.t. \; U, V \geq 0, \tag{1}$$

where $U = [u_1, ..., u_k] \in \mathbb{R}^{m \times k}$ is the basis matrix, $V = [v_1, ..., v_k] \in \mathbb{R}^{N \times k}$ is the coefficient matrix, and $\| \cdot \|_F$ is the Frobenius norm of a matrix. Usually, $k < \min(m, N)$, and thus we could consider $V$ as the low-dimensional representations for the training data $X$ with the objective of *best reconstruction* under non-negative constrains. However, the coefficient matrix derived based on the *best reconstruction* is unnecessarily good at discriminating power, since no label information is leveraged in NMF.

#### 2.1.2 Objective for Separability of Labeled Data

In order to reinforce the separability of the labeled data without the loss of construction capability, we divide the *reconstruction* representations $V$ into two parts, namely,

$$V = [V^1, V^2], \tag{2}$$

where $V^1 = [v_1^1, v_2^1, ..., v_q^1] \in \mathbb{R}^{N \times q}$ ($q < k$), which reserves the discriminative information for the labeled data. $V^2 = [v_1^2, v_2^2, ..., v_{k-q}^2] \in \mathbb{R}^{N \times (k-q)}$, which contains the additional reconstruction information together with $V^1$. Note that $V^1$ is expected to encode the major discriminative information, while the whole $V$ is used for data reconstruction purpose. Hence the targets of data reconstruction and classification coexist harmoniously, and do not mutually compromise as in conventional formulations with two objectives. Similarly, the basis matrix $U$ is also divided into two parts,

$$U = [U^1, U^2], \tag{3}$$

where $U^1 \in \mathbb{R}^{m \times q}$ and $U^2 \in \mathbb{R}^{m \times (k-q)}$.

There exist varieties of formulations for characterizing the separability of the labeled data, and Yan et al. (2007) claimed that most of them can be explained within a unified framework, called graph embedding. Let $G = \{X, S\}$ be an undirected weighted graph with vertex set $X$ and similarity matrix $S \in \mathbb{R}^{N \times N}$. Each element of $S$ measures for a pair of vertices the similarity, which is assumed to be non-negative in this work. The diagonal matrix $D$ and Laplacian matrix $L$ of a graph are defined as,

$$L = D - S, \;\; D_{ii} = \sum_{j \neq i} S_{ij}, \quad \forall\, i. \tag{4}$$

Graph embedding generally involves an intrinsic graph $G$, which characterizes the favorite relationship among the data, and a penalty graph $G^p = \{X, S^p\}$, which characterizes the unfavorable relationship among the data, with $L^p = D^p - S^p$, where $D^p$ is the diagonal matrix as defined in Eqn. (4). Then two targets of graph-preserving are given as follows,

$$\begin{cases} \max_{V^1} \sum_{i \neq j} \|V_i^1 - V_j^1\|^2 S_{ij}^p, \\ \min_{V^1} \; \sum_{i \neq j} \|V_i^1 - V_j^1\|^2 S_{ij}, \end{cases} \tag{5}$$

where $V_i^1$ is the $i$th rows of $V^1$. As aforementioned, $U^2$ is considered as the complementary space of $U^1$, and thus the first objective in Eqn. (5) can be approximately transformed into,

$$\min_{V^2} \sum_{i \neq j} \|V_i^2 - V_j^2\|^2 S_{ij}^p. \tag{6}$$

Note that $S_{ij}$ and $S_{ij}^p$ are set to zero in the above equations if either $x_i$ or $x_j$ is unlabeled.

#### 2.1.3 Objective for Regularization from both Labeled and Unlabeled Data

Compared with the labeled data which may require tedious human work, the unlabeled data are often much easier to obtain in real applications. The geometrical structure reflected by the interaction between the unlabeled data and labeled data could benefit the classification performance when the labeled data are not enough (Zhu, 2005). Here we adapt the *smoothness assumption*, which has been widely used in existing semi-supervised learning algorithms (Zhu, 2005), that nearby points in the original feature space tend to be close to each other in the new space and have similar class labels. The objective function for the regularization from both labeled and unlabeled data is given as,

$$\min_{V^1} \sum_{i \neq j} \|V_i^1 - V_j^1\|^2 S_{ij}^s, \tag{7}$$

where $S_{ij}^s$ could be defined based on the neighboring information as,

$$S_{ij}^s = \begin{cases} 1, & \text{if } x_i \in N_p(x_j) \text{ or } x_j \in N_p(x_i), \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $N_p(x_i)$ denotes the set of $p$ nearest neighbors of $x_i$. Similar to Eqn. (4), a diagonal matrix $D^s$ and a Laplacian matrix $L^s$ are also defined based on $S^s$. Note that both labeled and unlabeled data are used in Eqn. (7).

#### 2.1.4 Unified Objective Function

To achieve the above three objectives, we can have a unified objective function for N$^2$S$^2$L as,

$$\min_{U,V} \alpha(\sum_{i \neq j} \|V_i^1 - V_j^1\|^2 S_{ij} + \sum_{i \neq j} \|V_i^2 - V_j^2\|^2 S_{ij}^p) + \beta \sum_{i \neq j} \|V_i^1 - V_j^1\|^2 S_{ij}^s + \|X - UV^T\|_F^2, s.t.\ U, V \geq 0, \tag{9}$$

where $\alpha$ and $\beta$ are two positive parameters for balancing the aforementioned three objectives.

By simple algebraic deduction, Eqn. (9) can be rewritten as

$$\min_{U,V} \alpha Tr({V^1}^T L V^1) + \alpha Tr({V^2}^T L^p V^2) + \beta Tr({V^1}^T L^s V^1) + \|X - UV^T\|_F^2, s.t.\ U, V \geq 0. \tag{10}$$

Note that the above formulation is ill-posed, and the objective has the trend to drive $V$ to be zero. This issue is also suffered by the formulation for Fisher-NMF (Wang et al., 2004). As aforementioned, $U$ is the basis matrix and hence it is natural to require that the column vectors of $U$ are normalized, namely,

$$\|u_i\| = 1,\ i = 1, 2, \cdots, k. \tag{11}$$

This extra constraint makes the optimization problem more complicated, and in this work, we compensate the norms of the bases into the coefficient matrix and get the final object function for N$^2$S$^2$L as,

$$\min_{U,V} \|X - UV^T\|_F^2 + Tr[Q^1 {V^1}^T (\alpha L + \beta L^s) V^1 {Q^1}^T] + Tr[Q^2 {V^2}^T (\alpha L^p) V^2 {Q^2}^T], s.t.\ U, V \geq 0, \tag{12}$$

where

$$Q^1 = diag\{\|u_1\|, \|u_2\|, \cdots, \|u_q\|\}, \tag{13}$$
$$Q^2 = diag\{\|u_{q+1}\|, \|u_2\|, \cdots, \|u_k\|\}. \tag{14}$$

Note that as the matrices $S$, $S^p$, and $S^s$ are symmetric, thus the matrices $L$, $L^p$, and $L^s$ are also symmetric. This objective function is biquadratic, and generally there does not exist a closed-form solution. We present in the next subsection an iterative procedure for computing the nonnegative solution.

### 2.2 CONVERGENT ITERATIVE PROCEDURE

Most iterative procedures for solving high-order optimization problems transform the original intractable problem into a set of tractable sub-problems, and finally obtain the convergence to a local optimum. Our proposed iterative procedure also follows this philosophy and optimizes $U$ and $V$ alternately.

#### 2.2.1 Preliminaries

Before formally describing the iterative procedure for N$^2$S$^2$L, we first introduce the concept of auxiliary function, and the lemma which shall be used for the algorithmic deduction and convergence proof.

**Definition 1** Function $G(A, A')$ is an auxiliary function for function $F(A)$ if the following conditions are satisfied:

$$G(A, A') \geq F(A), \quad G(A, A) = F(A). \tag{15}$$

From the above definition, we have the following lemma with proof omitted.

**Lemma 1** If $G$ is an auxiliary function, then $F$ is nonincreasing under the updating rule

$$A^{t+1} = \arg\min_A G(A, A^t), \tag{16}$$

where $t$ means the $t$th iteration.

### 2.2.2 Optimize U for Given V

For a given $V$, the objective function in Eqn. (12) with respect to $U$ can be rewritten as

$$\begin{aligned} F(U) &= \|X - UV^T\|_F^2 \\ &+ Tr(Q^1 V^{1^T}(\alpha L + \beta L^s)V^1 Q^{1^T}) \\ &+ Tr(Q^2 V^{2^T}(\alpha L^p)V^2 Q^{2^T}) \\ &= \|X - UV^T\|_F^2 + Tr(UY_u U^T), \quad (17) \end{aligned}$$

where $Y_u$ is given as

$$\begin{aligned} Y_u &= \begin{bmatrix} V^{1^T}(\alpha L + \beta L^s)V^1 & 0 \\ 0 & V^{2^T}(\alpha L^p)V^2 \end{bmatrix} \cdot I \quad (18) \\ &= Y_{u+} - Y_{u-}, \quad (19) \end{aligned}$$

with the matrices $Y_{u+}$ and $Y_{u-}$ defined as,

$$\begin{aligned} Y_{u+} &= \begin{bmatrix} V^{1^T}(\alpha D + \beta D^s)V^1 & 0 \\ 0 & V^{2^T}(\alpha D^p)V^2 \end{bmatrix} \cdot I, \quad (20) \\ Y_{u-} &= \begin{bmatrix} V^{1^T}(\alpha S + \beta S^s)V^1 & 0 \\ 0 & V^{2^T}(\alpha S^p)V^2 \end{bmatrix} \cdot I. \quad (21) \end{aligned}$$

Here the operator $\cdot$ means that each element of the output matrix is the multiplication of the corresponding elements of two input matrices.

To integrate the non-negative constraints into the objective function, we set $\Upsilon_{ij}^u$ as the Lagrange multiplier for constraint $U_{ij} \geq 0$, and the matrix $\Upsilon^u = [\Upsilon_{ij}^u]$. Then the Lagrange $\mathcal{L}(U)$ with respect to $U$ is defined as,

$$\begin{aligned} \mathcal{L}(U) &= \|X - UV^T\|_F^2 + Tr(UY_u U^T) + Tr(\Upsilon^u U^T) \\ &= Tr(XX^T) - 2Tr(XVU^T) + Tr(UV^T VU^T) \\ &\quad + Tr(UY_u U^T) + Tr(\Upsilon^u U^T), \quad (22) \end{aligned}$$

By setting the derivation of $\mathcal{L}(U)$ with respect to $U$ as zero,

$$\frac{\partial \mathcal{L}(U)}{\partial U} = -2XV + 2UV^T V + 2UY_u + \Upsilon^u = 0, \quad (23)$$

along with the KKT condition (Kuhn & Tucker, 1951) of $\Upsilon_{ij}^u U_{ij} = 0$, we can have

$$\begin{aligned} & -(XV)_{ij}U_{ij} + (UV^T V)_{ij}U_{ij} + (UY_u)_{ij}U_{ij} \\ = & -(XV)_{ij}U_{ij} + (UV^T V)_{ij}U_{ij} \\ & +(UY_{u+})_{ij}U_{ij} - (UY_{u-})_{ij}U_{ij} \\ = & \ 0. \quad (24) \end{aligned}$$

Then for the final solution, the following relation should be satisfied,

$$U_{ij} \leftarrow U_{ij} \frac{(XV + UY_{u-})_{ij}}{(UV^T V + UY_{u+})_{ij}}. \quad (25)$$

We shall prove afterward that the above updating rule shall result in a convergent iterative procedure to obtain a local optimum solution. Obviously this updating rule is multiplicative and the non-negativity of the solution is guaranteed.

### 2.2.3 Convergence of the Updating Rule for U

Here, we denote $F_{ab}$ as the part of $F(U)$ relevant to $U_{ab}$, and then we have,

$$\begin{aligned} F'_{ab} &= (-2XV + 2UV^T V + 2UY_u)_{ab}, \quad (26) \\ F''_{ab} &= (2V^T V + 2Y_u)_{bb}. \quad (27) \end{aligned}$$

Then the auxiliary function of $F_{ab}$ is designed as

$$\begin{aligned} G(U_{ab}, U_{ab}^t) &= F_{ab}(U_{ab}^t) + F'_{ab}(U_{ab}^t)(U_{ab} - U_{ab}^t) \\ &+ \frac{(U^t V^T V)_{ab} + (U^t Y_{u+})_{ab}}{U_{ab}^t}(U_{ab} - U_{ab}^t)^2. \quad (28) \end{aligned}$$

**Lemma 2** Eqn. (28) is an auxiliary function for $F_{ab}$.

**Proof:** Since $G(U_{ab}, U_{ab}) = F_{ab}(U_{ab})$ is obvious, we need only show that $G(U_{ab}, U_{ab}^t) \geq F_{ab}(U_{ab})$. To do this, we compare the Taylor series expansion of $F_{ab}(U_{ab})$,

$$\begin{aligned} F_{ab}(U_{ab}) &= F_{ab}(U_{ab}^t) + F'_{ab}(U_{ab}^t)(U_{ab} - U_{ab}^t) \\ &+ \frac{1}{2}F''_{ab}(U_{ab} - U_{ab}^t)^2, \quad (29) \end{aligned}$$

with Eqn. (28), and then $G(U_{ab}, U_{ab}^t) \geq F_{ab}(U_{ab})$ is equivalent to

$$\frac{(U^t V^T V)_{ab} + (U^t Y_{u+})_{ab}}{U_{ab}^t} \geq (V^T V)_{bb} + (Y_u)_{bb}. \quad (30)$$

It is easy to verify that

$$(U^t V^T V)_{ab} = \sum_{m=1}^{k} U_{am}^t (V^T V)_{mb} \geq U_{ab}^t (V^T V)_{bb}, \quad (31)$$

and

$$\begin{aligned} (U^t Y_{u+})_{ab} &= \sum_{m=1}^{k} U_{am}^t (Y_{u+})_{mb} \geq U_{ab}^t (Y_{u+})_{bb} \\ &\geq U_{ab}^t (Y_{u+} - Y_{u-})_{bb} = U_{ab}^t (Y_u)_{bb}. \quad (32) \end{aligned}$$

Thus, Eqn. (30) holds and $G(U_{ab}, U_{ab}^t) \geq F_{ab}(U_{ab})$. □

**Lemma 3** Eqn. (25) could be obtained by minimizing the auxiliary function $G(U_{ab}, U_{ab}^t)$, where $U_{ab}^t$ is the iterative solution at the $t$th step.

**Proof:** To obtain the minimum, we only need set the derivative $\frac{\partial G(U_{ab}, U_{ab}^t)}{\partial U_{ab}} = 0$, and have

$$\begin{aligned} & \frac{\partial G(U_{ab}, U_{ab}^t)}{\partial U_{ab}} \\ = & \ F'_{ab}(U_{ab}^t) + \frac{2(U^t V^T V + U^t Y_{u+})_{ab}}{U_{ab}^t}(U_{ab} - U_{ab}^t) \\ = & \ 0. \quad (33) \end{aligned}$$

Then we can obtain the iterative updating rule for $U$ as,

$$U_{ij}^{t+1} \leftarrow U_{ij}^t \frac{(XV + U^t Y_{u-})_{ij}}{(U^t V^T V + U^t Y_{u+})_{ij}}, \quad (34)$$

and the lemma is proved. □

### 2.2.4 Optimize V for Given U

After updating the matrix $U$, we normalize the column vectors of $U$ and consequently convey the norm to the coefficient matrix $V$, namely,

$$\begin{aligned} u_m &\leftarrow u_m/\|u_m\|, \, \forall\, m, & (35)\\ v_m &\leftarrow v_m \times \|u_m\|, \, \forall\, m. & (36) \end{aligned}$$

Then based on the normalized $U$ in Eqn. (35), the objective function in Eqn. (12) with respect to $V$ is simplified to be,

$$\begin{aligned} F(V) &= \|X - UV^T\|_F^2 + Tr({V^1}^T(\alpha L + \beta L^s)V^1) \\ &\quad + Tr({V^2}^T(\alpha L^p)V^2) \\ &= \|X - UV^T\|_F^2 + Tr({V^1}^T Y_v^1 V^1) \\ &\quad + Tr({V^2}^T Y_v^2 V^2), \qquad (37) \end{aligned}$$

where $Y_v^1$ and $Y_v^2$ are given as,

$$\begin{aligned} Y_v^1 &= \alpha L + \beta L^s = Y_{v+}^1 - Y_{v-}^1, & (38)\\ Y_v^2 &= \alpha L^p = Y_{v+}^2 - Y_{v-}^2, & (39) \end{aligned}$$

with the matrices defined as,

$$\begin{aligned} Y_{v+}^1 &= \alpha D + \beta D^s, \quad Y_{v+}^2 = \alpha D^p, & (40)\\ Y_{v-}^1 &= \alpha S + \beta S^s, \quad Y_{v-}^2 = \alpha S^p. & (41) \end{aligned}$$

To integrate the non-negative constraints to the objective function, we set $\Upsilon_{ij}^v$ as the Lagrange multiplier for constraint $V_{ij} \geq 0$, and the matrix $\Upsilon^v = [\Upsilon_{ij}^v]$. Then the Lagrange $\mathcal{L}(V)$ with respect to $V$ is defined as,

$$\begin{aligned} \mathcal{L} &= \|X - UV^T\|_F^2 + Tr({V^1}^T Y_v^1 V^1) \\ &\quad + Tr({V^2}^T Y_v^2 V^2) + Tr(\Upsilon^v V^T) \\ &= Tr(XX^T) - 2Tr(XVU^T) \\ &\quad + Tr(UV^TVU^T) + Tr({V^1}^T Y_v^1 V^1) \\ &\quad + Tr({V^2}^T Y_v^2 V^2) + Tr(\Upsilon^v V^T). \qquad (42) \end{aligned}$$

By setting the derivation of $\mathcal{L}(V)$ with respect to $V$ as zero,

$$\frac{\partial \mathcal{L}(V)}{\partial V} = -2X^TU + 2VU^TU + 2[Y_v^1V^1, Y_v^2V^2] + \Upsilon^v = 0,$$

along with the KKT condition $\Upsilon_{ij}^v V_{ij} = 0$, we have

$$\begin{aligned} & -(X^TU)_{ij}V_{ij} + (VU^TU)_{ij}V_{ij} \\ & + [Y_v^1V^1, Y_v^2V^2]_{ij}V_{ij} \\ = \; & -(X^TU)_{ij}V_{ij} + (VU^TU)_{ij}V_{ij} \\ & + [Y_{v+}^1V^1, Y_{v+}^2V^2]_{ij}V_{ij} - [Y_{v-}^1V^1, Y_{v-}^2V^2]_{ij}V_{ij} \\ & = 0. \qquad (43) \end{aligned}$$

Then the following relation should be satisfied,

$$V_{ij} \leftarrow V_{ij}\frac{(X^TU + [Y_{v-}^1V^1, Y_{v-}^2V^2])_{ij}}{(VU^TU + [Y_{v+}^1V^1, Y_{v+}^2V^2])_{ij}}, \qquad (44)$$

which offers an updating rule for a convergent iterative procedure to obtain a local optimum solution for $V$.

### 2.2.5 Convergence of the Updating Rule for V

Here, we denote $F_{ab}$ as the part of $F(V)$ relevant to $V_{ab}$, and then we have,

$$F'_{ab} = (-2X^TU + 2VU^TU + 2[Y_v^1V^1, Y_v^2V^2])_{ab}, \quad (45)$$

$$F''_{ab} = \begin{cases} 2(U^TU)_{bb} + 2(Y_v^1)_{aa}, & \text{if } b \leq q, \\ 2(U^TU)_{bb} + 2(Y_v^2)_{aa}, & \text{otherwise.} \end{cases} \quad (46)$$

Then the auxiliary function of $F_{ab}$ is designed as,

$$\begin{aligned} G(V_{ab}, V_{ab}^t) &= F_{ab}(V_{ab}^t) + F'_{ab}(V_{ab}^t)(V_{ab} - V_{ab}^t) \\ &+ \frac{(V^tU^TU)_{ab} + [Y_{v+}^1{V^1}^t, Y_{v+}^2{V^2}^t]_{ab}}{V_{ab}^t}(V_{ab} - V_{ab}^t)^2. \quad (47) \end{aligned}$$

**Lemma 4** Eqn. (47) is an auxiliary function for $F_{ab}$.

**Proof:** Since $G(V_{ab}, V_{ab}) = F_{ab}(V_{ab})$ is obvious, we need only show that $G(V_{ab}, V_{ab}^t) \geq F_{ab}(V_{ab})$. To do this, we compare the Taylor series expansion of $F_{ab}(V_{ab})$

$$\begin{aligned} F_{ab}(V_{ab}) &= F_{ab}(V_{ab}^t) + F'_{ab}(V_{ab}^t)(V_{ab} - V_{ab}^t) \\ &\quad + \frac{1}{2}F''_{ab}(V_{ab} - V_{ab}^t)^2, \qquad (48) \end{aligned}$$

with Eqn. (47), and then $G(V_{ab}, V_{ab}^t) \geq F_{ab}(V_{ab})$ is equivalent to

$$\begin{aligned} & \frac{(V^tU^TU)_{ab} + [Y_{v+}^1{V^1}^t, Y_{v+}^2{V^2}^t]_{ab}}{V_{ab}^t} \\ & \geq \begin{cases} (U^TU)_{bb} + (Y_v^1)_{aa}, & \text{if } b \leq q, \\ (U^TU)_{bb} + (Y_v^2)_{aa}, & \text{otherwise.} \end{cases} \quad (49) \end{aligned}$$

It is easy to verify

$$(V^tU^TU)_{ab} = \sum_{m=1}^{k} V_{am}^t(U^TU)_{mb} \geq V_{ab}^t(U^TU)_{bb}, \quad (50)$$

and

$$\begin{aligned} & [Y_{v+}^1{V^1}^t, Y_{v+}^2{V^2}^t]_{ab} \\ = \; & \begin{cases} \sum_{m=1}^{N} (Y_{v+}^1)_{am}V_{mb}^t, & \text{if } b \leq q, \\ \sum_{m=1}^{N} (Y_{v+}^2)_{am}V_{mb}^t, & \text{otherwise.} \end{cases} \\ \geq \; & \begin{cases} (Y_{v+}^1)_{aa}V_{ab}^t, & \text{if } b \leq q, \\ (Y_{v+}^2)_{aa}V_{ab}^t, & \text{otherwise.} \end{cases} \\ \geq \; & \begin{cases} (Y_{v+}^1 - Y_{v-}^1)_{aa}V_{ab}^t, & \text{if } b \leq q, \\ (Y_{v+}^2 - Y_{v-}^2)_{aa}V_{ab}^t, & \text{otherwise.} \end{cases} \\ = \; & \begin{cases} (Y_v^1)_{aa}V_{ab}^t, & \text{if } b \leq q, \\ (Y_v^2)_{aa}V_{ab}^t, & \text{otherwise.} \end{cases} \qquad (51) \end{aligned}$$

Thus, Eqn. (49) holds and $G(V_{ab}, V_{ab}^t) \geq F_{ab}(V_{ab})$. □

**Lemma 5** Eqn. (44) could be obtained by minimizing the auxiliary function $G(V_{ab}, V_{ab}^t)$.

We omit the proof of Lemma 5 due to space limitation.

# 3 NON-NEGATIVE SEMI-SUPERVISED TENSOR FACTORIZATION

Tensor is a generalized concept of vector and matrix for data representation. Many research has shown that tensor-based data representation has the superiority over vector-based data representation for varieties of feature extraction algorithms, especially when the number of training data is small (Yan et al., 2007). In this section, we study the extension of the above vector-based non-negative semi-supervised learning algorithm to handle the tensor data of arbitrary order.

## 3.1 PROBLEM FORMULATION

Before formally introducing the tensor extension of N$^2$S$^2$L, we first redefine some notations. Let the training data $\mathcal{A} = [\mathcal{X}_1, ..., \mathcal{X}_N]$ be an $n$-th order tensor, in which each datum $\mathcal{X}_i \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_{n-1}}$ is represented as an ($n$-1)-th order tensor, *e.g.*, an image could be considered as a 2nd order tensor, namely matrix, and a video could be considered as a 3rd order tensor. Other notations are the same as in vector-based N$^2$S$^2$L.

For tensor data, we assume that the tensor $\mathcal{A}$ is factorized into the sum of $k$ rank-1 tensors as $\mathcal{A} = \sum_{m=1}^{k} (u_m^b \otimes)_{b=1}^{n-1} v_m$, and thus the objective function for tensorized non-negative semi-supervised learning is defined as,

$$\begin{aligned} \min_{U^b, V: 1 \leq b \leq n-1} \quad & \|\mathcal{A} - \sum_{m=1}^{k} (u_m^b \otimes)_{b=1}^{n-1} v_m\|_F^2 \\ & + Tr(Q^1 {V^1}^T (\alpha L + \beta L^s) V^1 {Q^1}^T) \\ & + Tr(Q^2 {V^2}^T (\alpha L^p) V^2 {Q^2}^T) \\ s.t. \quad & U^b, V \geq 0, \;\; 1 \leq b \leq n-1, \end{aligned} \tag{52}$$

where $\otimes$ is the outer product operator. The matrix $U^b = [u_1^b, ..., u_k^b] \in \mathbb{R}^{d_b \times k}$, $1 \leq b \leq n-1$, and $V = [v_1, ..., v_k] = [V^1, V^2]$. The matrices $Q^1$ and $Q^2$ are given by $Q^1 = \prod_{b=1}^{n-1} Q_b^1$ and $Q^2 = \prod_{b=1}^{n-1} Q_b^2$, where

$$\begin{aligned} Q_b^1 &= diag\{\|u_1^b\|, ..., \|u_q^b\|\}, \\ Q_b^2 &= diag\{\|u_{q+1}^b\|, ..., \|u_k^b\|\}. \end{aligned} \tag{53}$$

## 3.2 CONVERGENT ITERATIVE PROCEDURE

Similar to vector-based N$^2$S$^2$L, there does not exist a closed-form solution for Eqn. (52), and instead we propose to optimize the $V$ and $U^b$ iteratively. The optimization of $V$ is very similar to the vector-based N$^2$S$^2$L, and hence we omit the details here. Also for optimizing $U^b$, we only give the result here with the deduction details and convergence proof omitted.

The updating rule for $U^b$ for given $V$ and $U^p, p \neq b$ is,

$$U_{ij}^b \leftarrow U_{ij}^b \frac{(A_b Z_u + U^b Y_{u-})_{ij}}{(U^b Z_u^T Z_u + U^b Y_{u+})_{ij}}, \tag{54}$$

where $A_b$ is the matrix from the model-$b$ unfolding of the tensor $\mathcal{A}$, $Z_u$ is a matrix with its $m$th column defined as $[(u_m^p \otimes)_{p=b+1}^{n-1} (u_m^p \otimes)_{p=1}^{b-1} v_m]$, and

$$Y_{u+} = \begin{bmatrix} (\prod_{p \neq b} Q_p^1) {V^1}^T (\alpha D + \beta D^s) V^1 (\prod_{p \neq b} Q_p^1)^T & 0 \\ 0 & (\prod_{p \neq b} Q_p^2) {V^2}^T (\alpha D^p) V^2 (\prod_{p \neq b} Q_p^2)^T \end{bmatrix} \cdot I,$$

$$Y_{u-} = \begin{bmatrix} (\prod_{p \neq b} Q_p^1) {V^1}^T (\alpha S + \beta S^s) V^1 (\prod_{p \neq b} Q_p^1)^T & 0 \\ 0 & (\prod_{p \neq b} Q_p^2) {V^2}^T (\alpha S^p) V^2 (\prod_{p \neq b} Q_p^2)^T \end{bmatrix} \cdot I.$$

# 4 EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed non-negative semi-supervised learning (N$^2$S$^2$L) algorithm in three aspects: basis sparsity, discriminating power, and robustness to image occlusions. Due to the space limitation, we focus on the vector-based N$^2$S$^2$L only in all the experiments.

## 4.1 EXPERIMENT SETUP

Several popular subspace learning and semi-supervised learning algorithms are evaluated for comparison purpose: three unsupervised ones including principal component analysis (PCA) (Joliffe, 1986), non-negative matrix factorization (NMF) (Lee & Seung, 1999), and localized non-negative matrix factorization (LNMF) (Li et al., 2001), two supervised ones including linear discriminant analysis (LDA) (Belhumeur et al., 2002) and marginal fisher analysis (MFA) (Yan et al., 2007), one semi-supervised algorithm with feature dimension reduction, namely semi-supervised marginal fisher analysis (sMFA) (Yang et al., 2008b), three semi-supervised algorithms without feature dimension reduction including harmonic Gaussian field method (HGF) (Zhu et al., 2003), harmonic Gaussian field method coupled with the class mass normalization (HGF-CMN) (Zhu et al., 2003), and the consistency method (CONS) (Zhou et al., 2003).

For the N$^2$S$^2$L algorithm, the intrinsic graph and penalty graph are set as the same as those for MFA and sMFA, where the number of nearest neighbors of each sample is fixed as $n_c$-1 and the number of shortest pairs from different classes is set as 20 for each class in this work. For unsupervised and supervised algorithms, unlabeled data are only used for testing; while for semi-supervised algorithms, unlabeled data are used for both training and testing.

Two benchmark face database, *i.e.* ORL and FERET, are used. All images are aligned by fixing the locations of the two eyes. The ORL database contains 40 persons, each with 10 images. For the FERET database, we use 70 people with six images for each person. For ORL database, the images are normalized to 64-by-64 pixels; for FERET databases, the images are normalized to 56-by-46 pixels. For both databases, two images each person are randomly

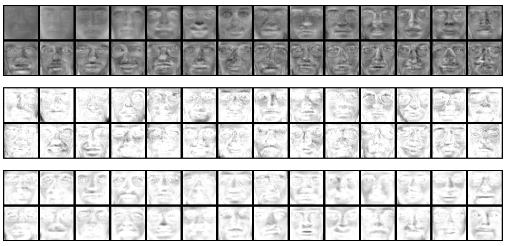

Figure 1: Basis matrix visualization of the algorithms PCA (1st row), NMF (2nd row), and N$^2$S$^2$L (3rd row) based on the training data of the ORL database.

selected as labeled data, while other images are considered as unlabeled data for testing. The performance is averaged over five random splits of labeled and unlabeled images.

### 4.2 SPARSITY ANALYSIS

In this subsection, we examine the sparsity property of the N$^2$S$^2$L algorithm. The basis matrices of N$^2$S$^2$L compared with those from PCA and NMF on ORL and FERET databases are depicted in Fig. 1 and Fig. 2, from which we can observe that the bases of N$^2$S$^2$L and NMF are much sparser than those of PCA. On the one hand, by leveraging labeled and unlabeled data, N$^2$S$^2$L may have superior discriminative capability over non-negative algorithms such as NMF and LNMF; on the other hand, the sparsity property of N$^2$S$^2$L makes it potentially more robust to image occlusions than PCA and other related algorithms do. We will validate these points in the next subsections.

### 4.3 CLASSIFICATION CAPABILITY

In this subsection, we evaluate the discriminating power of the N$^2$S$^2$L algorithm with five popular subspace learning algorithms: NMF, LNMF, PCA, LDA, and MFA, as well as four semi-supervised learning algorithms: sMFA, HGF, HGF-CMN, and CONS. For LDA and MFA, we first reduce the data to the dimension of $N_{tr}$-$N_c$ using PCA, where $N_{tr}$ is the number of labeled data and $N_c$ is the number of classes, for avoiding the singular value issue as conventionally. For sMFA, the data are reduced to the dimension of $N$-$N_c$, where $N$ is the number of all training images, including both labeled data and unlabeled data. For the non-negative algorithms NMF, LNMF and N$^2$S$^2$L, the parameter $k$ is set as $N_{tr} \times m/(N_{tr} + m)$ in all the experiment setting, and $q$ is simply set to be $N_c$ for N$^2$S$^2$L. The parameter $\beta$ in semi-supervised algorithms sMFA and N$^2$S$^2$L are selected from $[10^{-6}, 10^{-5}, ..., 10^3]$; while the parameter $\alpha$ in CONS are selected from $[0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99]$. The parameter $\alpha$ in N$^2$S$^2$L is simply set to be 10. For HGF, HGF-CMN, and CONS, we also use PCA preprocessing by retaining 90%, 95%, and 99% of the energy, in which the best results are used. We report the best results by exploring all possible feature dimensions for algorithms with dimensionality reduction as conventionally (Yan et al., 2007).

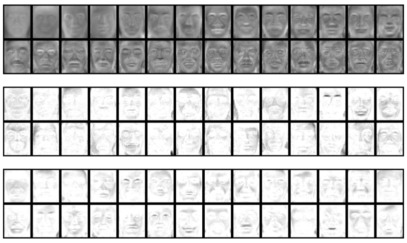

Figure 2: Basis matrix visualization of the algorithms PCA (1st row), NMF (2nd row), and N$^2$S$^2$L (3rd row) based on the training data of the FERET database.

Table 1: Face recognition accuracies (%) of different algorithms. Notice that the values in parentheses are the standard deviations of five rounds.

| Algorithm | ORL | FERET |
|---|---|---|
| NMF | 68.88 (±2.04) | 65.79 (±3.20) |
| LNMF | 70.69 (±2.26) | 73.36 (±3.50) |
| PCA | 70.88 (±1.49) | 69.71 (±3.33) |
| LDA | 75.38 (±3.83) | 77.88 (±2.16) |
| MFA | 76.50 (±3.41) | 77.79 (±3.25) |
| sMFA | **80.25 (±2.04)** | 80.64 (±3.25) |
| N$^2$S$^2$L | 79.19 (±2.08) | **80.71 (±3.45)** |
| HGF | 72.75 (±2.95) | 54.50 (±2.66) |
| HGF-CMN | 72.88 (±1.89) | 56.21 (±2.39) |
| CONS | 72.44 (±2.91) | 57.00 (±2.70) |

The comparison results of different algorithms on ORL and FERET databases are listed in Table 1, from which we could draw the following conclusions. First, the performances of non-negative algorithms NMF and LNMF are much worse than supervised algorithms LDA and MFA, and semi-supervised algorithms sMFA and N$^2$S$^2$L, which shows that without considering the labeled data, non-negative algorithms could not guarantee good discriminating power. Second, sMFA and N$^2$S$^2$L perform on average much better than LDA and MFA, which shows the importance of leveraging unlabeled data. Third, for semi-supervised algorithms without feature dimension reduction, HGF, HGF-CMN, and CONS, are consistently much worse than sMFA and N$^2$S$^2$L on these two face databases. One possible explanation for this may be that HGF, HGF-CMN, and CONS could only work well on densely sampled data sets, while for face recognition problem, there are too few images per person to reveal a meaningful manifold. Fourth, the performances of sMFA and N$^2$S$^2$L are comparable on average. It is reasonable since both of them fully utilize label and unlabeled data and the non-negative property are not necessary to have greater classification power. However, as shown in the next subsection, due to the sparsity property, N$^2$S$^2$L is much more robust than sMFA to image occlusions.

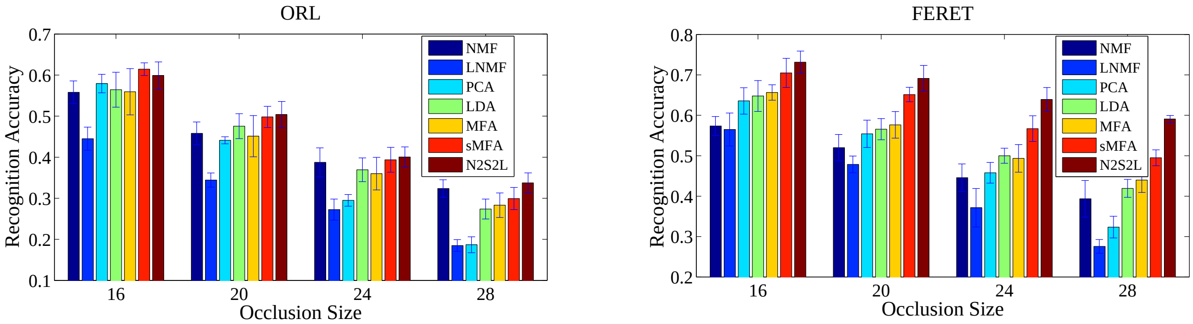

Figure 4: Face recognition accuracy vs. occlusion patch size. Left: results on the ORL face database. Right: results on the FERET face database. For better viewing, please see the color pdf file.

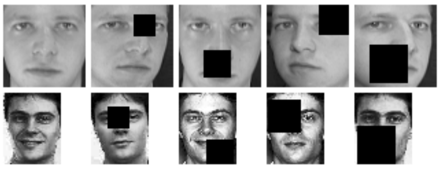

Figure 3: Sample images from ORL (up) and FERET (bottom) databases with occlusion patch sizes as 0-by-0, 16-by-16, 20-by-20, 24-by-24, and 28-by-28 pixels respectively.

### 4.4 ROBUSTNESS TO IMAGE OCCLUSIONS

As aforementioned, the bases from N$^2$S$^2$L are sparse, localized, and discriminative, which indicates that N$^2$S$^2$L is potentially more robust to image occlusions compared with other subspace learning and semi-supervised learning algorithms. To verify this point, we randomly add image occlusions of different sizes to the testing images (unlabeled images for semi-unsupervised algorithms). Notice that HGF, HGF-CMN, and CONS are transductive algorithms without feature dimension reduction, and hence we do not compare them here. Several example faces with occlusions of different sizes are depicted in Fig. 3. For each new datum, its coefficient vector is computed in the same way for NMF related algorithms as in (Li et al., 2001).

Fig. 4 shows the face recognition results of different algorithms. From these results, we can have the following observations: 1) sMFA and N$^2$S$^2$L still outperform other algorithms in most cases; and 2) for non-negative algorithms, NMF and N$^2$S$^2$L are more robust to image occlusions than other algorithms, more specifically, the gap between NMF and other algorithms becomes smaller, while the superiority of N$^2$S$^2$L over all other algorithms is more obvious as the occlusion patch size is bigger.

## References

Belhumeur, P., Hespanha, J., & Kriegman, D. (2002). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *TPAMI*, 711–720.

Belkin, M., Matveeva, I., & Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. *COLT*.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from examples. *JMLR*.

Cai, D., He, X., & Han, J. (2007). Semi-supervised discriminant analysis. *ICCV*.

Hazan, T., Polak, S., & Shashua, A. (2005). Sparse image coding using a 3d non-negative tensor factorization. *ICCV*, *1*, 50–57.

Joliffe, I. (1986). Principal component analysis. *Springer-Verlag, New York*.

Kotsia, I., Zafeiriou, S., & Pitas, I. (2007). A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *TIFS*, 588–595.

Kuhn, H., & Tucker, A. (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium*, 481–492.

Lee, D., & Seung, H. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*.

Li, S., Hou, X., Zhang, H., & Cheng, Q. (2001). Learning spatially localized, parts-based representation. *CVPR*.

Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. *ICML*.

Tao, H., Crabb, R., & Tang, F. (2005). Non-orthogonal binary subspace and its applications in computer vision. *CVPR*.

Wang, Y., Jiar, Y., Hu, C., & Turk, M. (2004). Fisher non-negative matrix factorization for learning local features. *ACCV*.

Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *TPAMI*.

Yang, J., Yan, S., Fu, Y., Li, X., & Huang, T. (2008a). Non-negative graph embedding. *CVPR*.

Yang, J., Yan, S., & Huang, T. (2008b). Ubiquitously supervised subspace learning. *TIP*.

Zhou, D., Bousquet, O., Lal, T., Weston, J., & Scholkopf, B. (2003). Learning with local and global consistency. *NIPS*.

Zhu, X. (2005). *Semi-supervised learning literature survey* (Technical Report Computer Sciences Technical Report 1530). University of Wisconsin-Madison.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *ICML*.