SAND2018-4770C

FULLY SUPERVISED NON-NEGATIVE MATRIX FACTORIZATION FOR FEATURE EXTRACTION

Woody Austin[a], *Dylan Anderson*[b], *Joydeep Ghosh*[c]

[a]University of Texas at Austin, Department of Computer Sciences
[b]Sandia National Laboratories, Albuquerque, NM
[c]University of Texas at Austin, Department of Electrical and Computer Engineering

ABSTRACT

Linear dimensionality reduction (DR) techniques have been applied with great success in the domain of hyperspectral image (HSI) classification. However, these methods do not take advantage of supervisory information. Instead, they act as a wholly unsupervised, disjoint portion of the classification pipeline, discarding valuable information that could improve classification accuracy. We propose the Supervised Non-negative Matrix Factorization (SNMF) to remedy this problem. By learning an NMF representation of the data jointly with a linear multi-class classifier, we are able to improve classification accuracy in real world problems. Experimental results on a widely used dataset show state of the art performance while maintaining full linearity of the entire DR and classification pipeline.

Index Terms— dimensionality reduction, non-negative matrix factorization, hyperspectral image classification

1. INTRODUCTION

Applications of Hyperspectral Image (HSI) classification are well known, ranging from precision agriculture [1] to disaster management [2]. Traditional techniques, such as support vector machines (SVM) [3] and k-nearest neighbors kNN [4], have been utilized for the HSI classification problem. These techniques can achieve reasonable performance, but in practice they are limited by several key properties of HSI. First, HSI pixels often contain hundreds of contiguous spectral bands, leading to high dimensional data. Second, measured spectra are often mixtures of a low number of unique components, yielding pixels that are highly co-linear. Third, the amount of labeled data available for model training and validation is often severely limited.

A common strategy to mitigate these properties is to first apply a dimensionality reduction (DR) transform to input spectra and then perform classification in the resulting low-dimensional space. Linear DR methods, such as principal component analysis (PCA) [5] and non-negative matrix factorization (NMF) [6], are widely used in HSI processing as they yield features that are interpretable and can often be assigned physical meaning. For instance, factors derived from NMF can be interpreted as endmembers and scores can be interpreted as the abundances of the endmembers within a pixel. However, performing DR prior to classification has a critical limitation: it makes no use of the supervisory information to find the desired low dimensional feature space. This is particularly problematic in settings where measured signal is dominated by background clutter, and direction of maximal variance does not align with the supervisory task. To address this shortcoming, we introduce supervised non-negative matrix factorization (SNMF) which fits a linear dimensionality reduction in the form of NMF *jointly* with learning classification weights for multinomial logistic regression.

Linear Discriminant Analysis (LDA) [7] is well known in the literature, but assumes rigid multi-variate normal structure over class distributions and does not impose non-negativity, limiting interpretation. Partial Least Squares (PLS) [8, 9] computes (effectively QR) decompositions of both training spectral data and labels. The projection matrices are found by maximizing joint covariance to encourage a relationship between data representation and classification. Other joint DR and classification formulations, such as Supervised Nonnegative Tensor Factorization with Maximum-Margin Constraint (SNTFM^2) [10] and Supervised Non-negative Tensor Factorization with Multinomial Logistic Regression (SNTFL) [11], have been developed in the literature. These formulations use the extracted factor matrix from NMF (or non-negative tensor decompositions) explicitly as input to the classifier. In contrast, the SNMF formulation proposed in this paper uses the learned spectral factor matrix to transform data into the feature space, allowing for test- and read-time predictions.

SNMF learns a linear operator that transforms raw data into a feature space that attempts to maximize classification performance. By performing DR and classification simultaneously, the learned feature space is biased towards the classification task thereby improving performance of the total pipeline. The subspace and classification boundary learned by SNMF are both linear, allowing for interpretability of the resulting model and reconstruction of projected data.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

2. METHOD

Non-Negative Matrix Factorization

Let $\mathbf{X} \in \mathbb{R}^{n_a \times n_b}$ be a pixels × wavelength hyperspectral image, where n_a denotes the number of pixels and n_b denotes the number of spectral bands. For a given spectral band, the 2-D image has been vectorized. Let $\mathbf{A} \in \mathbb{R}^{n_a \times k}$ and $\mathbf{B} \in \mathbb{R}^{n_b \times k}$ be an NMF approximation such that $\mathbf{X} \approx \mathbf{A}\mathbf{B}^\mathsf{T}$, with the entries of $\mathbf{A}$ and $\mathbf{B}$ are all greater than or equal to zero. The number of columns k in $\mathbf{A}$ and $\mathbf{B}$ is specified *apriori* and should reflect the rank of the data matrix; k can be interpreted as the number of end-members present in the scene. To find $\mathbf{A}$ and $\mathbf{B}$, we minimize the objective function:

$$\mathcal{L}_m = \frac{1}{2}||\mathbf{X} - \mathbf{A}\mathbf{B}^\mathsf{T}||_F^2 \tag{1}$$

subject to

$$\mathbf{A} \succeq \mathbf{0}, \quad \mathbf{B} \succeq \mathbf{0}. \tag{2}$$

This problem can be solved using the well-known Alternating Least Squares (ALS) algorithm [12].

SNMF

We consider the problem of performing a per-pixel classification. We denote the per-pixel class labels as $\mathbf{Y} \in \{0,1\}^{n_a \times n_c}$ where n_c is the number of classes present in the scene. Row i of $\mathbf{Y}$ is a one-hot encoded vector representing the class of pixel i. To incorporate the supervisory information $\mathbf{Y}$, we augment the NMF loss function given in Eq 1 as follows:

$$\mathcal{L} = \alpha\mathcal{L}_m + (1-\alpha)\mathcal{L}_s \tag{3}$$

$$\mathcal{L}_s = -\sum \mathbf{Y} * \log \hat{\mathbf{Y}}. \tag{4}$$

$$\hat{y}_{i,c} = \frac{e^{\mathbf{w}_c^\mathsf{T}(\mathbf{B}^\mathsf{T}\mathbf{x}_{i:})}}{\sum_{c'} e^{\mathbf{w}_{c'}^\mathsf{T}(\mathbf{B}^\mathsf{T}\mathbf{x}_{i:})}}. \tag{5}$$

where $*$ is the element-wise multiplication operation, and $\mathbf{W}$ is a parameter weight matrix of a linear multi-class logistic regression. Note that the additional term in the loss function (Eq 3) depends on the spectral-factor matrix $\mathbf{B}$, but not on the pixel-factor matrix $\mathbf{A}$, and the class decision boundary is a linear function of $\mathbf{W}$ and $\mathbf{B}$. Eq 5 differs from the standard formulation of the softmax function used in classification tasks because the raw data is transformed into a compressed feature space by $\mathbf{B}$. We intentionally avoid using the term projection because $\mathbf{B}$ is not a projection matrix; it only defines a linear transform into a feature space.

Incoming data will have the same number of raw features (n_b) as the original data matrix $\mathbf{X}$ and can be transformed to the compressed set of features (cardinality k) by applying $\mathbf{B}^\mathsf{T}$ to the data. The fit of the NMF decomposition becomes secondary to the classification task so that instead of focusing on matching the source data, we search for the best linear transformation of the data into a subspace that maximizes classification ability. Because it *is* a simple linear transformation, it is directly interpretable, computationally efficient at test time, and only requires a small memory footprint.

Other recent supervised DR methods from the literature [10, 11] have instead imposed supervision loss directly over the sample factor mode, $\mathbf{A}$, rather than $\mathbf{B}^\mathsf{T}\mathbf{X}$ as in SNMF. While more straightforward, this has a critical failing at model test time: the sample factor matrix corresponding to previously unseen data cannot be learned without the supervisory labels. By instead learning a supervised feature *extraction* rather than supervised features, SNMF allows for predictions over new data during inference.

The learning algorithm for SNMF is shown in Algorithm 1, and minimizes Eq 3 by alternating least squares. Since the supervisory term depends only on $\mathbf{W}$ and $\mathbf{B}$, updates for $\mathbf{A}$ proceed exactly as in unsupervised NMF. Solving for $\mathbf{W}$ follows the same procedure as finding weights in multinomial logistic regression given training data $\mathbf{B}^\mathsf{T}\mathbf{X}$ [13]. The gradient is given in Eq 6.

$$\frac{\partial \mathcal{L}_s}{\partial \mathbf{W}} = (\hat{\mathbf{Y}} - \mathbf{Y})(\mathbf{B}^\mathsf{T}\mathbf{X}) \tag{6}$$

The supervised loss gradient for $\mathbf{B}$ can be calculated analytically, shown in Eq 7.

$$\frac{\partial \mathcal{L}_s}{\partial \mathbf{B}} = \mathbf{X}(\hat{\mathbf{Y}} - \mathbf{Y})\mathbf{W}^\mathsf{T} \tag{7}$$

Note that the computation of $\frac{\partial \hat{\mathbf{Y}}}{\partial \mathbf{B}}$ in the chain rule is nontrivial and the individual rows, columns, or elements of $\mathbf{B}$ cannot be decoupled in the full objective function's derivative. This part of the gradient is included in Eq 7. It is important to note that in addition to updating $\hat{\mathbf{Y}}$ in Line 11, $\mathbf{B}$ and $\mathbf{W}$ are arguments to the construction of $\hat{\mathbf{Y}}$ and so it must be updated at each minimization step in Lines 8 and 10.

For incoming and unseen data $\mathbf{Z}$ at test time, we can form $\tilde{\mathbf{Z}} = \mathbf{B}^\mathsf{T}\mathbf{Z}$ as a new feature set. We then use the pre-trained multinomial logistic regression model to predict the class labels for $\tilde{\mathbf{Z}}$. The resulting classification pipeline is fully linear and uses only spectral information. For application spaces where data compression is desired, forming $\tilde{\mathbf{Z}}$ at read-time compresses the data by a factor of n_b/k where $k \ll n_b$.

Optimization details

We use scikit learn's `fmin_l_bfgs_b` solver to minimize Eqs 1 and 4 [14]. We employ optional $l1$ and $l2$ regularization for the NMF factors, adding the standard relevant terms to Eq 3 and its gradient. These regularization terms had minimal impact on accuracy, so we exclude them from further discussion and did not use them for the experiments in Section 3.

Algorithm 1 Supervised Non-negative Matrix Factorization (SNMF)

```
procedure SNMF($\mathbf{X}$, $\mathbf{Y}$, $k$, ftol)
    $\mathbf{A} \leftarrow$ Random initialization $\in \mathbb{R}^{n_a \times k}$
    $\mathbf{B} \leftarrow$ Random initialization $\in \mathbb{R}^{n_b \times k}$
    $\mathbf{W} \leftarrow$ Random initialization $\in \mathbb{R}^{k \times n_c}$
    $f_{new} \leftarrow 0$
    repeat
        $f_{old} \leftarrow f_{new}$
        $\mathbf{B} \leftarrow \mathrm{argmin}_{\mathbf{B}}(||\mathbf{X} - \mathbf{B}\mathbf{A}^{\mathsf{T}}||_F^2 - \sum \mathbf{Y} * \log \hat{\mathbf{Y}})$
        $\mathbf{A} \leftarrow \mathrm{argmin}_{\mathbf{A}} ||\mathbf{X} - \mathbf{B}\mathbf{A}^{\mathsf{T}}||_F^2$
        $\mathbf{W} \leftarrow \mathrm{argmin}_{\mathbf{W}}(-\sum \mathbf{Y} * \log \hat{\mathbf{Y}})$
        $\hat{\mathbf{Y}} \leftarrow \mathrm{softmax}(\mathbf{B}^{\mathsf{T}}\mathbf{X}, \mathbf{W})$
        $\mathcal{L}_m \leftarrow ||\mathbf{X} - \mathbf{B}\mathbf{A}^{\mathsf{T}}||_F^2$
        $\mathcal{L}_s \leftarrow -\sum \mathbf{Y} * \log \hat{\mathbf{Y}}$
        $f_{new} \leftarrow \alpha \mathcal{L}_m + (1 - \alpha)\mathcal{L}_s$
    until $|f_{old} - f_{new}| \leq$ ftol
    return $\mathbf{B}, \mathbf{A}, \mathbf{W}$
end procedure
```

The α parameter in Eq 3 corresponds to the relative weight of the decomposition to classification in the overall objective function. For HSI classification, decomposition fit is much less important than classification. Small but nonzero values of α yield the best performance, indicating that NMF serves as a structural regularization term to the feature transformation defined by $\mathbf{B}$.

3. RESULTS

We evaluate the proposed SNMF technique on the widely used Pavia Centre [15] HSI classification dataset. We remove pixels not considered to be one of the labeled classes from our analysis. Ten repeated trials of $50\% - 50\%$ train/test were performed, and overall accuracy (OA), average accuracy (AA) and κ-scores (κ) were computed. We compare SNMF against the following baselines. We use the Scikit-Learn version 0.19.0 [14] implementations of these baselines:

- **SVM:** Linear kernel support vector machine performed on full spectral data (no-DR).
- **PCA:** Principal component analysis followed by multinomial logistic regression for classification.
- **NMF:** Non-negative Matrix Factorization followed by multinomial logistic regression for classification.
- **LDA:** Multi-class Linear Discriminant Analysis [7].
- **PLS:** Partial Least Squares Discriminant Analysis [16]: extension of PLS to categorical data.

Overall accuracy results are shown in Figure 1 for a variety of DR ranks. Since SVM (full data size) and LDA ($n_c - 1 = 8$) do not parameterize rank, results are shown across all ranks for visual comparison. Supervised DR techniques (such as SNMF and PLS) encode supervisory information into their learned subspaces, providing more discriminative features and improved performance at low rank. In the case of the Pavia Centre dataset, SNMF achieves comparable performance with just 5 features. Best maximum overall performance is summarized in Table 1; for SNMF, PCA, PLS, and NMF, we use the rank yielding the best scoring model.

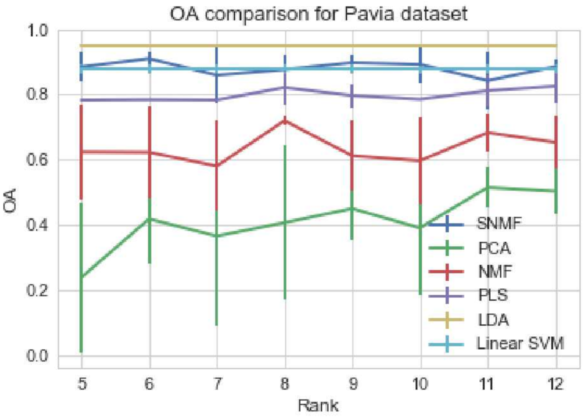

Fig. 1: OA vs. rank. Error bars represent 95% confidence intervals over 10 trials with 50/50 train/test split.

Method	Pavia Centre		
	OA	**AA**	κ
SNMF	**95.6%**	**87.1%**	**0.938**
PCA	72.2%	43.0%	0.597
NMF	77.3%	32.2%	0.649
PLS	89.1%	58.1%	0.842
LDA	95.3%	85.2%	0.934
SVM	91.1%	75.2%	0.874

Table 1: Comparison of the maximum Overall Accuracy, Average Accuracy, and κ scores for each method across all trials.

Since supervised methods learn more discriminative features, they require less training data. As shown in Figure 2, SNMF, PLS, and LDA perform well even with a small proportion of training data. In contrast, both PCA and NMF need much more data to reach optimal performance. This is a critical advantage for SNMF and other supervised DR techniques as the amount of labeled training data for HSI classification is often severely limited. Note that although SVM performs well with minimal data, it does not perform any DR. SNMF has the additional advantage of enforcing non-negativity in the derived features. Rank 8 was used for this study because it performed the best for this set of trials across baselines. This study was conducted with 10 randomized trials per point.

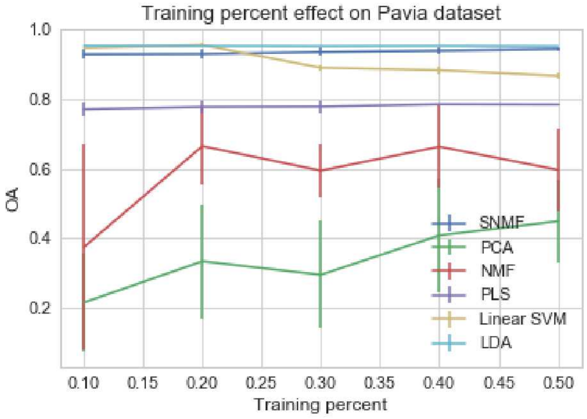

Fig. 2: OA vs fraction of data used for training.

4. CONCLUSIONS

This paper proposes Supervised Non-negative Matrix Factorization, an extension to the NMF model to encode supervisory information. SNMF jointly learns NMF and multinomial logistic regression. Joint learning of features and classification boundaries yields better features for downstream classification tasks, boosting performance. SNMF is shown experimentally to exhibit state-of-the-art performance for HSI classification and provides an advantage when the size of the learned subspace is small (e.g. for scenarios requiring significant read-time compression) or when the amount of labeled training data is limited. It exhibits these properties because its features directly encode the supervisory information, leading to a more compact and useful representation of the data for the downstream classification task. SNMF achieves this while remaining fully linear and non-negative, preserving interpretability of the learned feature spaces.

5. REFERENCES

[1] David J. Mulla, “Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps,” *Biosystems Engineering*, vol. 114, no. 4, pp. 358 – 371, 2013, Special Issue: Sensing Technologies for Sustainable Agriculture.

[2] Maya Nand Jha, Jason Levy, and Yang Gao, “Advances in remote sensing for oil spill disaster management: State-of-the-art sensors technology for oil spill surveillance,” *Sensors*, vol. 8, no. 1, pp. 236–255, 2008.

[3] Farid Melgani and Lorenzo Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[4] Li Ma, Melba M Crawford, and Jinwen Tian, “Local manifold learning-based k-nearest-neighbor for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4099–4109, 2010.

[5] H. Hotelling, “Analysis of a complex of statistical variables with principal components,” *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

[6] Daniel D Lee and H Sebastian Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems 13*, T K Leen, T G Dietterich, and V Tresp, Eds., pp. 556–562. MIT Press, 2001.

[7] R A Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, Sept. 1936.

[8] Abdi Herve, “Partial least squares (PLS) regression,” *Encyclopedia of Social Sciences Research Methods. Thousand Oaks (CA): Sage*, 2003.

[9] Svante Wold, Michael Sjöström, and Lennart Eriksson, “PLS-regression: a basic tool of chemometrics,” *Chemometrics Intellig. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, Oct. 2001.

[10] Fei Wu, Xu Tan, and Yi Yang, “Supervised nonnegative tensor factorization with Maximum-Margin constraint,” 2013, AAAI.

[11] Jingyun Choi et al., “Tensor-Factorization-Based phenotyping using group information: Case study on the efficacy of statins,” in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, New York, NY, USA, 2017, ACM-BCB ’17, pp. 516–525, ACM.

[12] H Kim and H Park, “Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method,” *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, Jan. 2008.

[13] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[14] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] Paolo Gambla, “Pavia hyperspectral dataset,” 2009.

[16] Matthew Barker and William Rayens, “Partial least squares for discrimination,” *J. Chemom.*, vol. 17, no. 3, pp. 166–173, Mar. 2003.