Thirty-seventh International Conference on Machine Learning

Adversarial Nonnegative Matrix Factorization

Lei Luo, Yanfu Zhang, Heng Huang
Electrical and Computer Engineering, University of Pittsburgh
JD Finance America Corporation
luoleipitt@gmail.com

Outline

- Background
- Motivation
- Our Work
- Experiments

Outline

- **Background**
- **Motivation**
- **Our Work**
- **Experiments**

Background

- The nonnegative matrix factorization (NMF) has been a prevalent nonnegative dimensionality reduction method
  - feature extraction, video tracking, image processing, and document clustering.
  - Popular models: standard NMF, RNMF(Truncated Cauchy NMF)
- What is the aim of nonnegative matrix factorization ?
  - It targets to factorize an $m \times N$-dimensional matrix $\mathbf{Y}$ into the product $\mathbf{AX}$ of two nonnegative matrices, with n columns in $\mathbf{A}$, where n is generally small.
- What make the success of nonnegative matrix factorization?
  - Successfully fitting noise term: $L_{2,1}$-norm loss Truncated Cauchy NMF loss
  - **Novel training approaches in model design**

Outline

- Background
- Motivation
- Our Work
- Experiments

Motivation

- The limitations of some existing methods
  - Existing methods are only suitable for some special types of noises, e.g., Laplacian or Cauchy noise, which cannot show the flexibility in facing the worst-case (i.e., adversarial) perturbations of data points.
- Our method
  - We introduce a novel Adversarial Nonnegative Matrix Factorization (ANMF) model by emphasizing potential test adversaries that are beyond the pre-defined constraints.

Outline

- Background
- Motivation
- Our Work
- Experiments

Our work

- NMF can be formulated as:

$$\min_{\mathbf{A},\mathbf{x}_1,\cdots,\mathbf{x}_N} \sum_{i=1}^{N} \| \mathbf{y}_i - \mathbf{A}\mathbf{x}_i \|_2^2, \quad (1)$$
$$s.t., \ \mathbf{A} \geq 0, \ \mathbf{x}_i \geq 0, i = 1, 2, \cdots, N.$$

Assumptions: *1. the learned feature data **A** and given data **Y** are drawn from an unknown distribution $\mathcal{D}$ at training time. The test data can be generated either from $\mathcal{D}$, the same distribution as the training data, or from $\tilde{\mathcal{D}}$, a modification of $\mathcal{D}$ generated by an attacker.*

2. The action of the learner is to select parameters $\{\mathbf{x}_1, \mathbf{x}_1, \cdots, \mathbf{x}_N\}$ of the Eq. (1). The attacker has an instance-specific target, and encourages that the prediction made by learner on the modified instance, , $\mathbf{y}_i = \tilde{\mathbf{A}}\mathbf{x}_i \ (i = 1, \cdots, N)$ is close to this target.

Our work

- The cost functions of each learner (Cl) and the attacker (Ca) are estimated by:

$$Cl(\mathbf{X}, \tilde{\mathbf{A}}, \mathbf{A}) = \alpha G(\tilde{\mathbf{A}}\mathbf{X}, \mathbf{Y}) + \beta G(\mathbf{A}\mathbf{X}, \mathbf{Y})$$

$$Ca(\mathbf{X}, \tilde{\mathbf{A}}, \mathbf{A}) = G(\tilde{\mathbf{A}}\mathbf{X}, \mathbf{Z}) + \lambda G(\tilde{\mathbf{A}}, \mathbf{A}).$$

- Ultimately, our model is expressed as:

$$\min_{\mathbf{X},\mathbf{A}} Cl(\mathbf{X}, \tilde{\mathbf{A}}^*(\mathbf{X}), \mathbf{A}) \; s.t., \; \tilde{\mathbf{A}}^*(\mathbf{X}) = \arg\min_{\tilde{\mathbf{A}}} Ca(\mathbf{X}, \tilde{\mathbf{A}}, \mathbf{A}),$$

$$\mathbf{X} \geq 0, \mathbf{A} \geq 0. \tag{2}$$

Theorem 1. Given **X**, the best response of the attacker is

$$\tilde{\mathbf{A}}^*(\mathbf{X}) = (\lambda\mathbf{A} + \mathbf{Z}\mathbf{X}^T)(\lambda\mathbf{I}_n + \mathbf{X}\mathbf{X}^T)^{-1}. \tag{3}$$

Since there is an inverse of complicated matrix in (3), it is difficult to solve problem (2) by directly substituting (3) into (2). To mitigate this limitation, we consider (3) as a constraint of (2), which leads to the following problem:

$$\min_{\mathbf{X},\mathbf{A},\tilde{\mathbf{A}}} Cl(\mathbf{X},\mathbf{A},\tilde{\mathbf{A}}),\ s.t.,\ \tilde{\mathbf{A}}(\lambda\mathbf{I}+\mathbf{X}\mathbf{X}^T)-(\lambda\mathbf{A}+\mathbf{Z}\mathbf{X}^T)=0, \quad \mathbf{X}\geq 0,\mathbf{A}\geq 0. \tag{4}$$

Let $\varphi(\tilde{\mathbf{A}},\mathbf{X})=\tilde{\mathbf{A}}(\lambda\mathbf{I}+\mathbf{X}\mathbf{X}^T)-(\lambda\mathbf{A}+\mathbf{Z}\mathbf{X}^T)$. Problem (4) can be approximated as:

$$\min_{\mathbf{X},\mathbf{A},\tilde{\mathbf{A}}} Cl(\mathbf{X},\mathbf{A},\tilde{\mathbf{A}})+\gamma\|\varphi(\tilde{\mathbf{A}},\mathbf{X})\|_F^2,\ s.t.,\mathbf{X}\geq 0,\mathbf{A}\geq 0. \tag{5}$$

Theoretical Analysis

We define the empirical reconstruction error of NMF as follows:

$$R_N(\mathbf{A}) = \frac{1}{N}\sum_{i=1}^{N} \min_{\mathbf{x}} ||\mathbf{y}_i - \mathbf{A}\mathbf{x}||_2^2.$$

Theorem 2. For ANMF problem, assume that $\mathbf{Y}$ is upper bound by 1. For any learned normalized $\mathbf{A}$ and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$|R(\mathbf{A}) - R_N(\mathbf{A})| \leq \min\left\{ \frac{14n\sqrt{n}}{\sqrt{N}} + \sqrt{\frac{n^2 \ln(16Nn)}{4N}} + \sqrt{\frac{\ln 2/\delta}{2N}}, \frac{2}{N} + \sqrt{\frac{mn\ln(4(1+n)\sqrt{m}nN) - \ln\frac{\delta}{2}}{2N}} \right\}.$$

Theorem 3. For orthogonal NMF problem, assume that $\mathbf{Y}$ upper bounded by 1. For any learned normalized $\mathbf{A}$ with and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$|R(\mathbf{A}) - R_N(\mathbf{A})| \leq 6n\sqrt{\frac{\pi}{N}} + \sqrt{\frac{\ln 2/\delta}{2N}}.$$

➢ **The proposed algorithm:** We apply the Alternating Direction Method of Multipliers (ADMM) optimization algorithm to solve our problem

Algorithm 1 Solving Eq. (9) via ADMM

Input: $\mathbf{Y} \in \mathbb{R}^{m \times N}$ and instance-specific target $\mathbf{Z}$

Output: feature matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and weight matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$.

Initialization: $\tilde{\mathbf{A}}$ and $\mathbf{X}$ using the traditional K-means method, $\mathbf{A}^0 = \mathrm{A}_K$, where $\mathbf{A}_K$ is the clustering centroid obtained by K-means method, $\mathbf{X}^0 = \mathbf{X}_K + 0.2$, where $\mathbf{X}_K$ is the K-means clustering result. $\mathbf{U}^0 = \mathbf{X}^{0T}$, $\tilde{\mathbf{B}}^0 = \tilde{\mathbf{A}}^0 = \mathbf{A}^0$, $\mathbf{M}^0 = \mathbf{0}$.

repeat

Update $(\mathbf{H}, \mathbf{B}, \mathbf{J})$ by

$$\mathbf{H}^{k+1} \leftarrow \mathbf{R}_1(2\alpha\mathbf{I}_N + \mu\mathbf{I}_N + 2\gamma\mathbf{U}^k\mathbf{U}^{kT})^{-1},$$

$$\tilde{\mathbf{B}}^{k+1} \leftarrow \max(0, \tilde{\mathbf{A}} + \frac{1}{\mu}\mathbf{M}_1^k),\ \mathbf{J}^{k+1} \leftarrow \frac{\mathbf{R}_2}{2\beta + \mu},$$

Update $(\mathbf{U}, \tilde{\mathbf{B}})$ by

$$\mathbf{U}^{k+1} \leftarrow \max(0, (2\gamma(\mathbf{H}^{k+1} - \mathbf{Z})^T(\mathbf{H}^{k+1} - \mathbf{Z}) + \mu\mathbf{I}_N)^{-1}\mathbf{R}_3),$$

$$\tilde{\mathbf{B}}^{k+1} \leftarrow \max(0, \mathbf{A}^k + \frac{1}{\mu}\mathbf{M}_1^k),$$

Update $\mathbf{A}$ by

$$\mathbf{A}^{k+1} \leftarrow \mathbf{R}_4(2\gamma\lambda^2\mathbf{I}_n + \mu\mathbf{I}_n + \mathbf{X}^k\mathbf{X}^{kT})^{-1},$$

Update $\tilde{\mathbf{A}}$ by

$$\tilde{\mathbf{A}}^{k+1} \leftarrow \mathbf{R}_5(2\gamma\lambda^2\mathbf{I}_n + \mu\mathbf{I}_n + \mathbf{X}^k\mathbf{X}^{kT})^{-1},$$

Update $\mathbf{X}$ by

$$\mathbf{X}^{k+1} \leftarrow ((\tilde{\mathbf{A}}^{k+1})^T\tilde{\mathbf{A}}^{k+1} + \mathbf{A}^{k+1T}\mathbf{A}^{k+1} + \mathbf{I}_n)^{-1}\mathbf{R}_6;$$

Update $\mathbf{M}$ by

$$\mathbf{M}^{k+1} \leftarrow \mathbf{M}^k + \mu(\phi(\tilde{\mathbf{A}}^{k+1}, \mathbf{A}^{k+1})\psi(\mathbf{X}^{k+1}) - \nu(\mathbf{H}^{\tilde{k+1}}, \mathbf{J}^{k+1}, \tilde{\mathbf{B}}^{k+1}, \mathbf{B}^{k+1}, \mathbf{U}^{k+1})).$$

until Converge

Our work

Convergence Analysis: To simplify notations, let us define

$$\Omega = (\mathbf{H}, \mathbf{B}, \mathbf{J}, \mathbf{U}, \tilde{\mathbf{B}}, \mathbf{A}, \tilde{\mathbf{A}}, \mathbf{X}, \mathbf{M}).$$

Theorem 4. Let $\{\Omega_k\}_{k=1}^{\infty}$ be a sequence generated by Algorithm 1 that satisfies the condition

$$\lim_{k\to\infty} (\Omega^{k+1} - \Omega^k) = 0.$$

Then any accumulation point of $(\mathbf{A}^k, \tilde{\mathbf{A}}^k, \mathbf{X}^k)_k^{\infty}$ is a KKT point of problem (5).

Outline

- Background
- Motivation
- Our Work
- Experiments

Experiments

Table 2. ACC of noise-free Real Datasets. The best results are marked in bold.

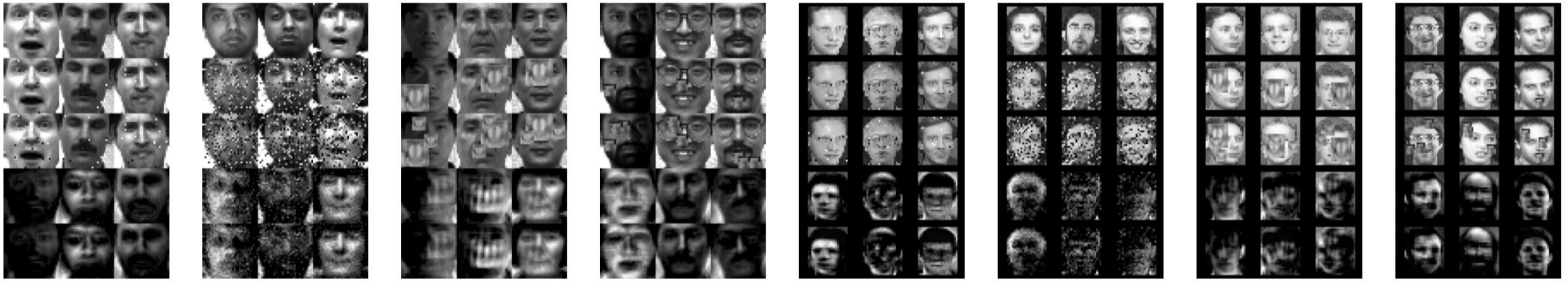
Dataset	NMF	ONMF	$L2$, 1-norm NMF	CNMF	ANMF
MNIST	$0.7933(\pm 0.0497)$	$0.7987(\pm 0.0441)$	$0.8027(\pm 0.0410)$	$0.7947(\pm 0.0228)$	$\mathbf{0.8067}(\pm 0.0490)$
Yale	$0.4388(\pm 0.0233)$	$0.4145(\pm 0.0360)$	$0.4424(\pm 0.0235)$	$0.4036(\pm 0.0380)$	$\mathbf{0.4509}(\pm 0.0164)$
ORL	$0.7005(\pm 0.0060)$	$0.6420(\pm 0.0356)$	$0.6895(\pm 0.0139)$	$0.5935(\pm 0.0323)$	$\mathbf{0.7305}(\pm 0.0294)$
UMIST	$0.4880(\pm 0.0285)$	$0.4616(\pm 0.0295)$	$0.4845(\pm 0.0255)$	$0.4442(\pm 0.0235)$	$\mathbf{0.4946}(\pm 0.0186)$
COIL-20	$0.6692(\pm 0.0215)$	$0.6626(\pm 0.0264)$	$0.6578(\pm 0.0130)$	$0.6601(\pm 0.0300)$	$\mathbf{0.6833}(\pm 0.0162)$
USPS	$0.7468(\pm 0.0004)$	$0.7738(\pm 0.0003)$	$0.7550(\pm 0.0002)$	$0.7429(\pm 0.0050)$	$\mathbf{0.7780}(\pm 0.0002)$
BBCSport	$0.9493(\pm 0.0007)$	$0.9460(\pm 0.0024)$	$0.9468(\pm 0.0006)$	$0.9327(\pm 0.0064)$	$\mathbf{0.9531}(\pm 0.0031)$
BBC	$0.9604(\pm 0.0011)$	$0.9619(\pm 0.0028)$	$0.9597(\pm 0.0002)$	$0.9202(\pm 0.0032)$	$\mathbf{0.9649}(\pm 0.0010)$
WebKB	$0.6619(\pm 0.0095)$	$0.6657(\pm 0.0038)$	$0.6618(\pm 0.0083)$	$0.6525(\pm 0.0117)$	$\mathbf{0.6672}(\pm 0.0084)$
Reuters	$0.7836(\pm 0.0059)$	$0.7495(\pm 0.0164)$	$0.7788(\pm 0.0071)$	$0.7197(\pm 0.0112)$	$\mathbf{0.8047}(\pm 0.0098)$
RCV	$0.6458(\pm 0.0194)$	$0.6493(\pm 0.0054)$	$0.6420(\pm 0.0183)$	$0.6280(\pm 0.0021)$	$\mathbf{0.6516}(\pm 0.0137)$
TDT2	$0.8546(\pm 0.0067)$	$0.8246(\pm 0.0119)$	$0.8448(\pm 0.0046)$	$0.8062(\pm 0.0150)$	$\mathbf{0.8638}(\pm 0.0176)$

Experiments

Table 3. ACC of noisy Real Datasets. The best results are marked in bold.

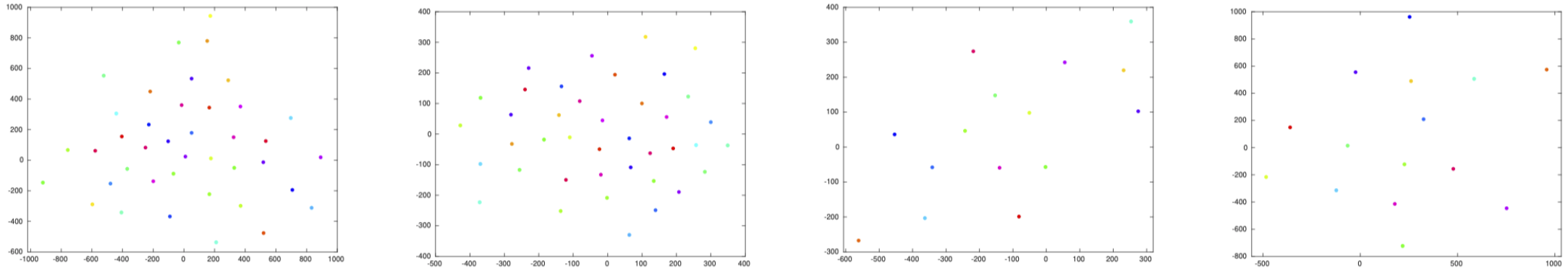
Noise	Dataset	NMF	ONMF	$L2$, 1-norm NMF	CNMF	ANMF
SP	MNIST	0.8067(±0.0464)	0.8093(±0.0379)	0.8080(±0.0477)	0.8067(±0.0254)	**0.8160**(±0.0421)
	Yale	0.3879(±0.0321)	0.3527(±0.0248)	0.3806(±0.0168)	0.3576(±0.0223)	**0.4036**(±0.0180)
	UMIST	0.4800(±0.0150)	0.4602(±0.0268)	0.4814(±0.0086)	0.4275(±0.0162)	**0.5078**(±0.0124)
	ORL	0.6155(±0.0252)	0.5475(±0.0083)	0.6225(±0.0275)	0.5350(±0.0173)	**0.6670**(±0.0248)
	COIL-20	0.6723(±0.0247)	0.6547(±0.0190)	0.6762(±0.0175)	0.6782(±0.0316)	**0.6830**(±0.0194)
	USPS	0.7542(±0.0004)	0.7716(±0.0003)	0.7592(±0.0003)	0.7505(±0.0058)	**0.7793**(±0.0002)
Pixel	MNIST	0.7880(±0.0417)	0.7733(±0.0194)	0.7800(±0.0481)	0.7453(±0.0202)	**0.8027**(±0.0293)
	Yale	0.3867(±0.0301)	0.3261(±0.0464)	0.3576(±0.0424)	0.3394(±0.0346)	**0.4012**(±0.0286)
	UMIST	0.4706(±0.0261)	0.4483(±0.0170)	0.4720(±0.0235)	0.4310(±0.0269)	**0.4866**(±0.0223)
	ORL	0.5370(±0.0141)	0.4850(±0.0157)	0.5145(±0.0192)	0.4650(±0.0190)	**0.5600**(±0.0366)
	COIL-20	0.6829(±0.0117)	0.6469(±0.0209)	0.6850(±0.0293)	0.6229(±0.0345)	**0.6924**(±0.0337)
	USPS	0.7520(±0.0002)	0.7638(±0.0009)	0.7527(±0.0007)	0.7269(±0.0003)	**0.7654**(±0.0006)
regular	MNIST	0.8107(±0.0494)	0.8040(±0.0376)	0.8093(±0.0543)	0.7920(±0.0311)	**0.8160**(±0.0423)
	Yale	0.3597(±0.0175)	0.3547(±0.0309)	0.3651(±0.0330)	0.3519(±0.0158)	**0.3852**(±0.0273)
	UMIST	0.4525(±0.0302)	0.4737(±0.0242)	0.4710(±0.0259)	0.4223(±0.0248)	**0.4828**(±0.0251)
	ORL	0.5465(±0.0243)	0.5145(±0.0288)	0.5520(±0.0198)	0.4695(±0.0368)	**0.5680**(±0.0207)
	COIL-20	0.5274(±0.0209)	0.5145(±0.0121)	0.5278(±0.0054)	0.5293(±0.0284)	**0.5315**(±0.0198)
	USPS	0.5210(±0.0005)	0.5296(±0.0074)	0.5195(±0.0018)	0.5306(±0.0078)	**0.5327**(±0.0065)
irregular	MNIST	0.2493(±0.0037)	0.2440(±0.0060)	0.2427(±0.0060)	0.2480(±0.0056)	**0.2497**(±0.0163)
	Yale	0.5468(±0.0261)	0.4982(±0.0458)	0.5406(±0.0266)	0.4861(±0.0262)	**0.5549**(±0.0301)
	UMIST	0.2247(±0.0056)	0.2115(±0.0089)	0.2235(±0.0087)	0.2136(±0.0051)	**0.2271**(±0.0057)
	ORL	0.3250(±0.0127)	0.2880(±0.0132)	0.3230(±0.0110)	0.2780(±0.0082)	**0.3485**(±0.0146)
	COIL-20	0.6792(±0.0202)	0.6706(±0.0228)	0.6782(±0.0181)	0.6586(±0.0177)	**0.6827**(±0.0145)
	USPS	0.7388(±0.0002)	**0.7602**(±0.0001)	0.7455(±0.0001)	0.7320(±0.0001)	0.7559(±0.0003)

Experiments



(a) S & P (b) pixel (c) regular (d) irregular (e) S & P (f) pixel (g) regular (h) irregular

Figure 1. Illustrations of face datasets Yale (a)-(d) and ORL (e)-(h) with different types of noises (pixel, regular, irregular). From top row to bottom row: origin, noisy data, noisy $\mathbf{Z}$, $\mathbf{A}$, $\tilde{\mathbf{A}}$



(a) $\tilde{\mathbf{A}}$ on ORL (b) $\mathbf{A}$ on ORL (c) $\tilde{\mathbf{A}}$ on Yale (d) $\mathbf{A}$ on Yale

Figure 2. Visualizing Feature Matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ via T-SNE on ORL and Yale Datasets

References

- Guan, N., Liu, T., Zhang, Y., Tao, D., and Davis, L. S. Truncated cauchy non-negative matrix factorization for robust subspace learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
- Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. arXiv preprint arXiv:1811.07457, 2018.
- Hajinezhad, D., Chang, T.-H., Wang, X., Shi, Q., and Hong, M. Nonnegative matrix factorization using admm: Algorithm and convergence analysis. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 4742–4746. IEEE, 2016.
- ……

Thank you