Supplementary Material for: Adversarial Nonnegative Matrix Factorization

Abstract

This supplementary document contains all the technical proofs and several experimental results for the ICML'20 paper entitled "Adversarial Nonnegative Matrix Factorization". It is actually the appendix section of the paper. The technical proofs are provided in Appendix A. All experiments are detailed in Appendix B.

1 Appendix A

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

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

Proof. We can derive the best response of the attacker by using the first order condition. □

Lemma A1. [S1] For NMF problems, assume that each $\mathbf{y}_i$ $(i = 1, 2, \cdots, N)$ is upper bounded by 1. For any learned normalized **A** and any $\delta > 0$ with probability at least $1 - \delta$, we have

$$|R(\mathbf{A}) - R_N(\mathbf{A})| \leq \frac{14\sqrt{n}}{\sqrt{N}} + \sqrt{\frac{r^2 \ln(16Nn)}{4N}} + \sqrt{\frac{\ln\frac{2}{\delta}}{2N}}.$$

Lemma A2. [S2] For NMF problems, assume that each $\mathbf{y}_i$ $(i = 1, 2, \cdots, N)$ is upper bounded by 1. For any learned normalized **A** and any $\delta > 0$ with probability at least $1 - \delta$, we have

$$|R(\mathbf{A}) - R_N(\mathbf{A})| \leq \frac{2}{N} + \sqrt{\frac{mn \ln(4(1+n)\sqrt{mn}N) - \ln\frac{\delta}{2}}{2N}}.$$

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

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

Proof. The regularization in model (10) shrinks the search space for optimizing the bases **A**. According the definitions of the Rademacher complexity and covering number, we know that he induced Rademacher complexity and covering number of ANMF are smaller that those NMF. Then, Connecting Lemma A1 Lemma A2, Theorem 1 can be easily proved. □

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

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

Proof. According to Proposition 3.2 in [S1], we can know that the included Rademacher complexity for NMF is upper bounded by

$$\frac{\sqrt{2\pi}}{n}\left\{\sqrt{8}\sup_{\mathbf{A}\in\mathbb{R}_{+}^{m\times n}}\sum_{i=1}^{N}\sum_{k=1}^{n}g_{ik}\langle\mathbf{y}_i,\mathbf{Ae}_k\rangle+\sqrt{2}E\sup_{\mathbf{A}\in\mathbb{R}_{+}^{m\times n}}\sum_{i=1}^{N}\sum_{l,k=1}^{n}g_{ilk}\langle\mathbf{A}e_l,\mathbf{Ae}_k\rangle\right\}$$

where g_{ik} and g_{ilk} are orthogonal Gaussian sequences that are of independent Gaussian random variables with zero mean and unit standard deviation.

Thus, for orthogonal NMF, we have

$$\begin{aligned}\mathfrak{R}(F)&\leq\frac{\sqrt{2\pi}}{N}\left(\sqrt{8}\sup_{\mathbf{A}\in\mathbb{R}_{+}^{m\times n}}\sum_{i=1}^{N}\sum_{k=1}^{n}g_{ik}\langle\mathbf{y}_i,\mathbf{Ae}_k\rangle+\sqrt{2}E\sup_{\mathbf{A}\in\mathbb{R}_{+}^{m\times n}}\sum_{i=1}^{N}\sum_{l,k=1}^{n}g_{ilk}\langle\mathbf{Ae}_l,\mathbf{Ae}_k\rangle\right)\\&\leq\frac{\sqrt{2\pi}}{N}\left(\sqrt{8}\sup_{\mathbf{A}\in\mathbb{R}_{+}^{m\times n}}\sum_{i=1}^{N}\sum_{k=1}^{n}g_{ik}\langle\mathbf{y}_i\mathbf{Ae}_k\rangle+\sqrt{2}E\sup_{\mathbf{A}\in\mathbb{R}_{+}^{m\times n}}\sum_{i=1}^{N}\sum_{l,k=1}^{n}g_{ik}\parallel\mathbf{Ae}_k\parallel_2^2\right)\\&\leq\frac{\sqrt{2\pi}}{N}(n\sqrt{8N}+n\sqrt{2N})\\&=6n\sqrt{\frac{\pi}{N}}\end{aligned}$$

where the second inequality holds, because the columns of **A** are orthogonal, and the third inequality holds because of Lemma 3.3 in [S1]. Connecting Theorem 1 of [S2], we can complete the proof.

Theorem 4. The optimal solution $\mathbf{X}^*$ of the problem in (6) can be considered as an approximate solution to the optimization problem (16).

Proof. The proof is similar to [26, Theorem 6]. Here, we omit the detailed proof process.

The derivation of Algorithm 1:

ADMM is applied to minimizing the augmented Lagrangian problem (19) with respect to $\mathbf{H},\mathbf{J},\mathbf{U},\mathbf{X},\mathbf{A},\tilde{\mathbf{A}},\mathbf{B},\tilde{\mathbf{B}}$ alternately. The iterative scheme of ADMM for problem (18) is given as follows:

$$(\mathbf{H}^{k+1},\mathbf{B}^{k+1},\mathbf{J}^{k+1})=\underset{\mathbf{H},\mathbf{B}\geq0,\mathbf{J}}{\text{argmin}}\,L_\mu(\mathbf{H},\mathbf{B},\mathbf{J},\mathbf{U}^k,\tilde{\mathbf{B}}^k,\mathbf{A}^k,\tilde{\mathbf{A}}^k,\mathbf{X}^k,\mathbf{M}^k);\tag{A.1}$$

$$(\mathbf{U}^{k+1},\tilde{\mathbf{B}}^{k+1})=\underset{\mathbf{U}\geq0,\tilde{\mathbf{B}}\geq0}{\text{argmin}}\,L_\mu(\mathbf{H}^{k+1},\mathbf{B}^{k+1},\mathbf{J}^{k+1},\mathbf{U},\tilde{\mathbf{B}},\mathbf{A}^k,\tilde{\mathbf{A}}^k,\mathbf{X}^k,\mathbf{M}^k);\tag{A.2}$$

$$\mathbf{A}^{k+1}=\underset{\mathbf{A}}{\text{argmin}}\,L_\mu(\mathbf{H}^{k+1},\mathbf{B}^{k+1},\mathbf{J}^{k+1},\mathbf{U}^{k+1},\tilde{\mathbf{B}}^{k+1},\mathbf{A},\tilde{\mathbf{A}}^k,\mathbf{X}^k,\mathbf{M}^k);\tag{A.3}$$

$$\tilde{\mathbf{A}}^{k+1}=\underset{\tilde{\mathbf{A}}}{\text{argmin}}\,L_\mu(\mathbf{H}^{k+1},\mathbf{B}^{k+1},\mathbf{J}^{k+1},\mathbf{U}^{k+1},\tilde{\mathbf{B}}^{k+1},\mathbf{A}^{k+1},\tilde{\mathbf{A}},\mathbf{X}^k,\mathbf{M}^k);\tag{A.4}$$

$$\mathbf{X}^{k+1}=\underset{\mathbf{X}}{\text{argmin}}\,L_\mu(\mathbf{H}^{k+1},\mathbf{B}^{k+1},\mathbf{J}^{k+1},\mathbf{U}^{k+1},\tilde{\mathbf{B}}^{k+1},\mathbf{A}^{k+1},\tilde{\mathbf{A}}^{k+1},\mathbf{X},\mathbf{M}^k);\tag{A.5}$$

$$\mathbf{M}^{k+1}=\mathbf{M}^{k+1}+\mu(\phi(\tilde{\mathbf{A}}^{k+1},\mathbf{A}^{k+1})\psi(\mathbf{X}^{k+1})-\nu(\mathbf{H}^{\widetilde{k+1}},\mathbf{J}^{k+1},\tilde{\mathbf{B}}^{k+1},\mathbf{B}^{k+1},\mathbf{U}^{k+1})).\tag{A.6}$$

Here

$$\begin{gathered}\phi(\tilde{\mathbf{A}},\mathbf{A})=\text{diag}(\tilde{\mathbf{A}},\mathbf{A},\tilde{\mathbf{A}},\mathbf{A},\mathbf{I}_N),\psi(\mathbf{X})=\text{diag}(\mathbf{X},\mathbf{X},\mathbf{I}_n,\mathbf{I}_n,\mathbf{X}^T),\\\nu(\tilde{\mathbf{H}},\mathbf{J},\tilde{\mathbf{B}},\mathbf{B},\mathbf{U})=\text{diag}(\mathbf{H},\mathbf{J},\tilde{\mathbf{B}},\mathbf{B},\mathbf{U}),\mathbf{M}=\text{diag}(\mathbf{M}_1,\mathbf{M}_2,\mathbf{M}_3,\mathbf{M}_4,\mathbf{M}_5).\end{gathered}\tag{A.7}$$

(I) The (**H**, **B**, **J**)-subproblem, namely problem (A.1), can be written as the following three separate problems:

$$\min_{\mathbf{H}} \alpha\|\mathbf{H}-\mathbf{Y}\|_F^2+\gamma\|\mathbf{H}\mathbf{U}^k+\lambda\tilde{\mathbf{A}}^k-\lambda\mathbf{A}^k-\mathbf{Z}\mathbf{U}^k\|_F^2+\frac{\mu}{2}\|\tilde{\mathbf{A}}^k\mathbf{X}^k-\mathbf{H}+\frac{1}{\mu}\mathbf{M}_3^k\|_F^2; \tag{A.8}$$

$$\min_{\mathbf{B}\geq 0} \|\tilde{\mathbf{A}}^k-\tilde{\mathbf{B}}+\frac{1}{\mu}\mathbf{M}_1^k\|_F^2; \tag{A.9}$$

$$\min_{\mathbf{J}} \beta\|\mathbf{J}\ -\mathbf{Y}\|_F^2+\frac{\mu}{2}\|\tilde{\mathbf{A}}^k\mathbf{X}^k-\mathbf{J}+\frac{1}{\mu}\mathbf{M}_3^k\|_F^2. \tag{A.10}$$

Observe that problems (A.8, (A.9) and (A.10) are convex quadratic problem, hence the optimal solution can be achieved by directly calculating the stationary points w.r.t. **H**, **B** and **J** for the corresponding objective function. Denote

$$\mathbf{R}_1=2\alpha\mathbf{Y}+2\gamma(\lambda\mathbf{A}^k+\mathbf{Z}\mathbf{U}^k-\lambda\tilde{\mathbf{A}}^k){\mathbf{U}^k}^T+\mu(\tilde{\mathbf{A}}^k\mathbf{X}^k+\frac{1}{\mu}\mathbf{M}_3^k),\ \mathbf{R}_2=2\beta\mathbf{Y}+\mu(\tilde{\mathbf{A}}^k\mathbf{X}^k+\frac{1}{\mu}\mathbf{M}_3^k), \tag{A.11}$$

we have

$$\mathbf{H}^{\ k+1}=\mathbf{R}_1(2\alpha\mathbf{I}_N+\mu\mathbf{I}_N+2\gamma\mathbf{U}^k{\mathbf{U}^k}^T)^{-1};\ \tilde{\mathbf{B}}^{\ k+1}=\max(0,\tilde{\mathbf{A}}+\frac{1}{\mu}\mathbf{M}_1^k);\ \mathbf{J}^{\ k+1}=\frac{\mathbf{R}_2}{2\beta+\mu}. \tag{A.12}$$

(II) Ignoring the constant terms of the object function in (A.2), $(\mathbf{U},\tilde{\mathbf{B}})$-sub-problem is acquired as follows:

$$\min_{\mathbf{U}\geq 0} \gamma\|\mathbf{H}^{k+1}\mathbf{U}+\lambda\tilde{\mathbf{A}}^k-\lambda\mathbf{A}^k-\mathbf{Z}\mathbf{U}\|_F^2+\frac{\mu}{2}\|{\mathbf{X}^k}^T-\mathbf{U}+\frac{1}{\mu}\mathbf{M}_5^{k+1}\|_F^2. \tag{A.13}$$

$$\min_{\tilde{\mathbf{B}}\geq 0} \|\tilde{\mathbf{A}}^k-\tilde{\mathbf{B}}^{k+1}+\frac{1}{\mu}\mathbf{M}_1^k\|_F^2. \tag{A.14}$$

then we get immediately the closed-form solution of $(\mathbf{U},\tilde{\mathbf{B}})$-sub-problem:

$$\mathbf{U}^{k+1}=\max(0,(2\gamma(\mathbf{H}^{k+1}-\mathbf{Z})^T(\mathbf{H}^{k+1}-\mathbf{Z})+\mu\mathbf{I}_N)^{-1}\mathbf{R}_3),\ \tilde{\mathbf{B}}^{k+1}=\max(0,\mathbf{A}^k+\frac{1}{\mu}\mathbf{M}_1^k), \tag{A.15}$$

where

$$\mathbf{R}_3=2\gamma(\mathbf{H}^{k+1}-\mathbf{Z})^T(\lambda\mathbf{A}^k-\lambda\tilde{\mathbf{A}}^k)+\mu({\mathbf{X}^k}^T+\frac{1}{\mu}\mathbf{M}_5^k). \tag{A.16}$$

Similarly, for sub-problems (A.3), (A.4) and (A.5), we can achieve their closed-form solutions as:

$$\mathbf{A}^{k+1}=\mathbf{R}_4(2\gamma\lambda^2\mathbf{I}_n+\mu\mathbf{I}_n+\mathbf{X}^k{\mathbf{X}^k}^T)^{-1}, \tag{A.17}$$

$$\tilde{\mathbf{A}}^{k+1}=\mathbf{R}_5(2\gamma\lambda^2\mathbf{I}_n+\mu\mathbf{I}_n+\mathbf{X}^k{\mathbf{X}^k}^T)^{-1}, \tag{A.18}$$

$$\mathbf{X}^{k+1}=((\tilde{\mathbf{A}}^{k+1})^T\tilde{\mathbf{A}}^{k+1}+{\mathbf{A}^{k+1}}^T\mathbf{A}^{k+1}+\mathbf{I}_n)^{-1}\mathbf{R}_6, \tag{A.19}$$

where

$$\mathbf{R}_4=2\gamma\lambda(\mathbf{H}^{k+1}\mathbf{U}^{k+1}+\lambda\tilde{\mathbf{A}}^k-\mathbf{Z}\mathbf{U}^{k+1})+\mu(\mathbf{B}^{k+1}-\frac{1}{\mu}\mathbf{M}_2^k)+\mu(\mathbf{J}^{k+1}-\frac{1}{\mu}\mathbf{M}_4^k){\mathbf{X}^k}^T, \tag{A.20}$$

$$\mathbf{R}_5=2\gamma\lambda(\lambda\mathbf{A}^{k+1}+\mathbf{Z}\mathbf{U}^{k+1}-\mathbf{H}^{k+1}\mathbf{U}^{k+1})+\mu(\tilde{\mathbf{B}}^{k+1}-\frac{1}{\mu}\mathbf{M}_1^k)+\mu(\mathbf{H}^{k+1}-\frac{1}{\mu}\mathbf{M}_3^k){\mathbf{X}^k}^T, \tag{A.21}$$

$$\mathbf{R}_6=(\tilde{\mathbf{A}}^{k+1})^T(\mathbf{H}^{k+1}-\frac{1}{\mu}\mathbf{M}_3^k)+(\mathbf{A}^{k+1})^T(\mathbf{J}^{k+1}-\frac{1}{\mu}\mathbf{M}_4^k)+{\mathbf{U}^{k+1}}^T-\frac{1}{\mu}{\mathbf{M}_5^k}^T. \tag{A.22}$$

Summarizing the above analysis, the detailed process for solving problem (10) is presented in Algorithm 1.

To derive the KKT conditions for problem (17), we first write the Lagrangian function of (17) as follows:

$$\begin{aligned} L(\mathbf{H},\mathbf{B},\mathbf{J},\mathbf{U},\tilde{\mathbf{B}},\mathbf{A},\tilde{\mathbf{A}},\mathbf{X},\mathbf{M}) = {} & \alpha\|\mathbf{H}-\mathbf{Y}\|_F^2 + \beta\|\mathbf{J}\ -\mathbf{Y}\|_F^2 + \gamma\|\mathbf{H}\mathbf{U} + \lambda\tilde{\mathbf{A}} - \lambda\mathbf{A} - \mathbf{Z}\mathbf{U}\|_F^2 \\ & + Tr(\mathbf{M}_1^T(\tilde{\mathbf{A}} - \tilde{\mathbf{B}})) + Tr(\mathbf{M}_2^T(\mathbf{A}-\mathbf{B})) + Tr(\mathbf{M}_3^T(\tilde{\mathbf{A}}\mathbf{X} - \mathbf{H})) \\ & + Tr(\mathbf{M}_4^T(\mathbf{A}\mathbf{X}-\mathbf{J})) + Tr(\mathbf{M}_5^T(\mathbf{X}^T - \mathbf{U})). \end{aligned} \tag{A.23}$$

Therefore, a point Ω is a KKT point of problem (17) if it satisfies the KKT conditions for problem (17):

$$2\alpha(\mathbf{H}-\ \mathbf{Y}) + 2\gamma(\mathbf{H}\mathbf{U} + 2\gamma\lambda\tilde{\mathbf{A}} - \lambda\mathbf{A} - \mathbf{Z}\mathbf{U})\mathbf{U}^T - \mathbf{M}_3 = 0, \tag{A.24}$$

$$2\lambda\gamma(\mathbf{H}\mathbf{U} + \lambda\tilde{\mathbf{A}} - \lambda\mathbf{A} - \mathbf{Z}\mathbf{U}) + \mathbf{M}_1 +\ \mathbf{M}_3\mathbf{X}^T = 0, \tag{A.25}$$

$$2\lambda\gamma(\lambda\mathbf{A} + \mathbf{Z}\mathbf{U} - \mathbf{H}\mathbf{U} - \lambda\tilde{\mathbf{A}}) + \mathbf{M}_2 +\ \mathbf{M}_4\mathbf{X}^T = 0, \tag{A.26}$$

$$2\gamma(\mathbf{H}-\mathbf{U})^T(\mathbf{H}\mathbf{U} + \lambda\tilde{\mathbf{A}} - \lambda\mathbf{A} - \mathbf{Z}\mathbf{U}) + \mathbf{M}_5 = 0, \tag{A.27}$$

$$\tilde{\mathbf{A}}^T\mathbf{M}_3 + \mathbf{A}^T\mathbf{M}_4 + \mathbf{M}_5^T = 0, \tag{A.28}$$

$$\phi(\tilde{\mathbf{A}},\mathbf{A})\psi(\mathbf{X}) - \nu(\tilde{\mathbf{H}},\mathbf{J},\tilde{\mathbf{B}},\mathbf{B},\mathbf{U}) = 0, \tag{A.29}$$

$$\mathbf{M}_1 \leq 0 \leq \tilde{\mathbf{B}}, Tr(\mathbf{M}_1^T\tilde{\mathbf{B}}) = 0, \mathbf{M}_2 \leq 0 \leq \mathbf{B}, Tr(\mathbf{M}_2^T\mathbf{B}) = 0, \mathbf{M}_5 \leq 0 \leq \mathbf{U}, Tr(\mathbf{M}_5^T\mathbf{U}) = 0. \tag{A.30}$$

Theorem 5. Let $\{\Omega_k\}_{k=1}^{\infty}$ be a sequence generated by Algorithm 1 that satisfies the condition

$$\lim_{k\to\infty}(\Omega^{k+1} - \Omega^k) = 0. \tag{A.31}$$

Then any accumulation point of $\{\Omega^k\}_{k=1}^{\infty}$ is a KKT point of problem (17). Consequently, any accumulation point of $(\mathbf{A}^k, \tilde{\mathbf{A}}^k, \mathbf{X}^k)_k^{\infty}$ is a KKT point of problem (9).

proof. We can rearrange the update formulas w.r.t. $\mathbf{H}$ in Algorithm 1 into:

$$(\mathbf{H}_+ - \mathbf{H})(2\alpha\mathbf{I}_N + \mu\mathbf{I}_N + 2\gamma\mathbf{U}^k{\mathbf{U}^k}^T) = -(2\alpha(\mathbf{H}-\ \mathbf{Y}) + 2\gamma(\mathbf{H}\mathbf{U} + 2\gamma\lambda\tilde{\mathbf{A}} - \lambda\mathbf{A} - \mathbf{Z}\mathbf{U})\mathbf{U}^T - \mathbf{M}_3 + \tilde{\mathbf{A}}\mathbf{X} - \mathbf{H}), \tag{A.32}$$

The assumption $\Omega_+ - \Omega \to 0$ implies that the left-and right-hand sides above all go to zero. Now we add subscript *k* to all variables $\mathbf{H}, \mathbf{A}, \tilde{\mathbf{A}}, \mathbf{X}, \mathbf{U}$, and replacing $\mathbf{H}_+$ by $\mathbf{H}_{k+1}$. Letting *k* go to infinity and noting $\mathbf{H}_{k+1} = \mathbf{H}_k + (\mathbf{H}_{k+1} - \mathbf{H}_k)$, where the second term vanishes asymptotically, we have

$$2\alpha(\mathbf{H}^k - \mathbf{Y}) + 2\gamma(\mathbf{H}^k\mathbf{U}^k + 2\gamma\lambda\tilde{\mathbf{A}}^k - \lambda\mathbf{A}^k - \mathbf{Z}\mathbf{U}^k{\mathbf{U}^k}^T - \mathbf{M}_3^k \to 0, \tag{A.33}$$

Similarly, we have

$$2\lambda\gamma(\mathbf{H}^k\mathbf{U}^k + \lambda\tilde{\mathbf{A}}^k - \lambda\mathbf{A}^k - \mathbf{Z}\mathbf{U})^k + \mathbf{M}_1^k +\ \mathbf{M}_3^k{\mathbf{X}^k}^T \to 0, \tag{A.34}$$

$$2\lambda\gamma(\lambda\mathbf{A}^k + \mathbf{Z}\mathbf{U}^k - \mathbf{H}^k\mathbf{U}^k - \lambda\tilde{\mathbf{A}}^k) + \mathbf{M}_2^k +\ \mathbf{M}_4^k{\mathbf{X}^k}^T \to 0, \tag{A.35}$$

$$2\gamma(\mathbf{H}^k - \mathbf{U}^k)^T(\mathbf{H}^k\mathbf{U}^k + \lambda\tilde{\mathbf{A}}^k - \lambda\mathbf{A}^k - \mathbf{Z}\mathbf{U}^k) + \mathbf{M}_5^k \to 0, \tag{A.36}$$

$$(\tilde{\mathbf{A}}^k)^T\mathbf{M}_3^k + (\mathbf{A}^k)^T\mathbf{M}_4^k + {\mathbf{M}_5^k}^T \to 0, \tag{A.37}$$

$$\phi(\tilde{\mathbf{A}}^k, \mathbf{A}^k)\psi(\mathbf{X}^k) - \nu(\tilde{\mathbf{H}}^k, \mathbf{J}^k, \tilde{\mathbf{B}}^k, \mathbf{B}^k, \mathbf{U}^k) \to 0, \tag{A.38}$$

$$Tr((\mathbf{M}_1^k)^T \tilde{\mathbf{B}}^k) \to 0, \tag{A.39}$$

$$Tr((\mathbf{M}_2^k)^T \mathbf{B}^k) \to 0, \tag{A.40}$$

$$Tr((\mathbf{M}_5^k)^T \mathbf{U}^k) \to 0. \tag{A.41}$$

Clearly, the first six equations in the KKT conditions (A.33)-(A.38) for problem (17) are satisfied at any limit point

$$\Omega^* = (\mathbf{H}^*, \mathbf{B}^*, \mathbf{J}^*, \mathbf{U}^*, \tilde{\mathbf{B}}^*, \mathbf{A}^*, \tilde{\mathbf{A}}^*, \mathbf{X}^*, \mathbf{M}^*). \tag{A.42}$$

In addition, the non-positivity of $\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_5\}$ also can be easily verified using the nonnegativity of **B**, $\tilde{\mathbf{B}}$ and **U**. This completes the proof. □

2 Appendix B

In this section we include extensive experiments, including the comparison with baselines on various real-life data, and some illustration demonstrating the characterization of the proposed method.

This section is organized as follow: § 2.1 describes the involved data; § 2.2 establishes the experimental setting; § 2.3 and § 2.4 compares the proposed method with state-of-the-art baselines; § 2.5 demonstrates the convergence of the proposed ADMM solver.

2.1 Data Description

The detailed information about the datasets are summarized in Table 1. These datasets come from real-life, and enjoys a wide dynamic range of class, dimensions and sample sizes.[1]

Table 1: Description of Benchmark Datasets

Dataset	Number of Instances	Dimensions	Classes	Category
MNIST	150	784	10	image
Yale	165	1024	15	image
ORL	400	644	40	image
UMIST	575	644	20	image
COIL-20	1440	1024	20	image
USPS	9298	256	10	image
BBCsports	737	4613	5	text
BBCNews	2225	9635	5	text
WebKB	4199	7770	4	text
Reuters	9298	256	10	text
RCV	9625	29992	4	text
TDT2	9394	36771	30	text

2.2 Baseline, Pre-processing and Evaluation Strategy

Some representative methods, including, Standard Nonnegative Matrix Factorization (SNMF), $L_{2,1}$-norm based NMF model [19], Orthogonal Nonnegative Matrix Factorization (ONMF) [4], and Capped norm Nonnegative Matrix Factorization (CNMF) [9], are compared with the ANMF. It should noted that the main novelty of this paper is to consider potential test adversaries in modeling, not the robust characterization for noise. Thus, it is unfair to compare our method with some robust methods such as Correntropy induced NMF model and Truncated CauchyNMF model. Throughout the experiments, we set ANMF parameters as $\alpha = 0.6$, $\beta = 10^{-5}$, $\gamma = 10^{-3}$, $\lambda = 10^{-3}$, and $\mu = 1$.

[1] The data involved here comes from `http://www.cad.zju.edu.cn/home/dengcai/Data/data.html`.

Each data point of the image dataset is normalized as a vector with unit length. For the document dataset, the TFIDF term weight normalization is applied to each data point and the TFIDF vector is also normalized. For each dataset, the value n (*i.e.*, the number of the columns of feature matrix $\mathbf{A}$) is set as the real number of classes. We run K-Means method 50 times and choose the best clustering result, corresponding to the lowest objective value, to initialize all baselines and the proposed methods. The popular Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) are chosen as our metric of performance. We report the mean and standard error based on five runs of experiments.

2.3 Experimental Results on Noise-free Data

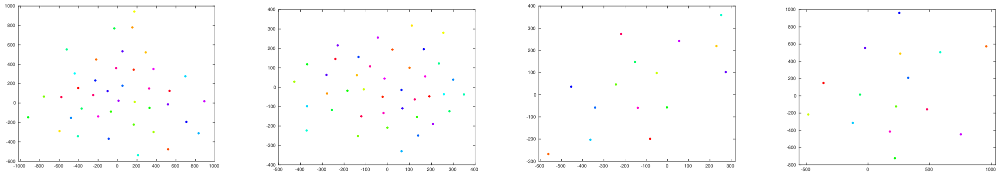
The detailed results for clustering accuracy and normalized mutual information results are shown in Table 2 and Table 3 (The best results are marked in bold). It can be observed that the advantage of ANMF is quite evident. Although $L_{2,1}$-norm based NMF is a robust NMF method, the ignoring of test adversaries leads to the undesired performance. Compared to NMF, ONMF achieves the better results, which indicates that the orthogonal constraint *w.r.t.* weight matrix $\mathbf{X}$ is helpful for improving the performance of models. However, our method is more competitive than other methods on the all databases. Therefore, considering the adversarial perturbations in modeling can increase the robustness of NMF. Figure 1 illustrates distribution of $\mathbf{A}$ and $\tilde{\mathbf{A}}$ with dimensionality reduced using T-SNEWe can see that all data points are fully separated.

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

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	$0.7800(\pm 0.0340)$	$0.7933(\pm 0.0497)$	$0.7987(\pm 0.0441)$	$0.8027(\pm 0.0410)$	$0.7947(\pm 0.0228)$	**0.8067**(± 0.0490)
Yale	$0.4109(\pm 0.0410)$	$0.4388(\pm 0.0233)$	$0.4145(\pm 0.0360)$	$0.4424(\pm 0.0235)$	$0.4036(\pm 0.0380)$	**0.4509**(± 0.0164)
ORL	$0.7035(\pm 0.0297)$	$0.7005(\pm 0.0060)$	$0.6420(\pm 0.0356)$	$0.6895(\pm 0.0139)$	$0.5935(\pm 0.0323)$	**0.7305**(± 0.0294)
UMIST	$0.4797(\pm 0.0267)$	$0.4880(\pm 0.0285)$	$0.4616(\pm 0.0295)$	$0.4845(\pm 0.0255)$	$0.4442(\pm 0.0235)$	**0.4946**(± 0.0186)
COIL-20	$0.6629(\pm 0.0244)$	$0.6692(\pm 0.0215)$	$0.6626(\pm 0.0264)$	$0.6578(\pm 0.0130)$	$0.6601(\pm 0.0300)$	**0.6833**(± 0.0162)
USPS	$0.7706(\pm 0.0005)$	$0.7468(\pm 0.0004)$	$0.7738(\pm 0.0003)$	$0.7550(\pm 0.0002)$	$0.7429(\pm 0.0050)$	**0.7780**(± 0.0002)
BBCSport	$0.9463(\pm 0.0077)$	$0.9493(\pm 0.0007)$	$0.9460(\pm 0.0024)$	$0.9468(\pm 0.0006)$	$0.9327(\pm 0.0064)$	**0.9531**(± 0.0031)
BBC	$0.9633(\pm 0.0026)$	$0.9604(\pm 0.0011)$	$0.9619(\pm 0.0028)$	$0.9597(\pm 0.0002)$	$0.9202(\pm 0.0032)$	**0.9649**(± 0.0010)
WebKB	$0.6603(\pm 0.0036)$	$0.6619(\pm 0.0095)$	$0.6657(\pm 0.0038)$	$0.6618(\pm 0.0083)$	$0.6525(\pm 0.0117)$	**0.6672**(± 0.0084)
Reuters	$0.8010(\pm 0.0174)$	$0.7836(\pm 0.0059)$	$0.7495(\pm 0.0164)$	$0.7788(\pm 0.0071)$	$0.7197(\pm 0.0112)$	**0.8047**(± 0.0098)
RCV	$0.6447(\pm 0.0109)$	$0.6458(\pm 0.0194)$	$0.6493(\pm 0.0054)$	$0.6420(\pm 0.0183)$	$0.6280(\pm 0.0021)$	**0.6516**(± 0.0137)
TDT2	$0.8629(\pm 0.0145)$	$0.8546(\pm 0.0067)$	$0.8246(\pm 0.0119)$	$0.8448(\pm 0.0046)$	$0.8062(\pm 0.0150)$	**0.8638**(± 0.0176)

Table 3: NMI of noise-free Real Datasets. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	$0.7429(\pm 0.0197)$	$0.7581(\pm 0.0318)$	$0.7580(\pm 0.0225)$	$0.7572(\pm 0.0334)$	$0.7439(\pm 0.0166)$	**0.7677**(± 0.0278)
Yale	$0.4571(\pm 0.0436)$	$0.4827(\pm 0.0224)$	$0.4591(\pm 0.0367)$	$0.4870(\pm 0.0191)$	$0.4482(\pm 0.0336)$	**0.4940**(± 0.0226)
ORL	$0.8377(\pm 0.0231)$	$0.8339(\pm 0.0075)$	$0.8029(\pm 0.0211)$	$0.8343(\pm 0.0146)$	$0.7651(\pm 0.0325)$	**0.8471**(± 0.0177)
UMIST	$0.5998(\pm 0.0146)$	$0.6011(\pm 0.0146)$	$0.5813(\pm 0.0220)$	$0.5970(\pm 0.0109)$	$0.5578(\pm 0.0250)$	**0.6120**(± 0.0099)
COIL-20	$0.7554(\pm 0.0169)$	$0.7547(\pm 0.0117)$	$0.7507(\pm 0.0170)$	$0.7553(\pm 0.0082)$	$0.7397(\pm 0.0188)$	**0.7581**(± 0.0160)
USPS	$0.6655(\pm 0.0006)$	$0.6365(\pm 0.0004)$	$0.6620(\pm 0.0005)$	$0.6437(\pm 0.0002)$	$0.6320(\pm 0.0032)$	**0.6665**(± 0.0004)
BBCSport	$0.8601(\pm 0.0158)$	$0.8579(\pm 0.0021)$	$0.8560(\pm 0.0064)$	$0.8522(\pm 0.0017)$	$0.8303(\pm 0.0089)$	**0.8690**(± 0.0103)
BBC	$0.8856(\pm 0.0054)$	$0.8773(\pm 0.0030)$	$0.8806(\pm 0.0059)$	$0.8760(\pm 0.0006)$	$0.7999(\pm 0.0065)$	**0.8875**(± 0.0022)
WebKB	$0.3577(\pm 0.0130)$	$0.3637(\pm 0.0070)$	$0.3645(\pm 0.0101)$	$0.3638(\pm 0.0070)$	$0.3532(\pm 0.0066)$	**0.3689**(± 0.0013)
Reuters	$0.4885(\pm 0.0147)$	$0.4661(\pm 0.0074)$	$0.4397(\pm 0.0123)$	$0.4613(\pm 0.0041)$	$0.4005(\pm 0.0089)$	**0.4862**(± 0.0073)
RCV	$0.0002(\pm 0.0000)$	**0.0003**(± 0.0000)	**0.0003**(± 0.0000)	$0.0002(\pm 0.0000)$	**0.0003**(± 0.0000)	**0.0003**(± 0.0000)
TDT2	$0.7643(\pm 0.0257)$	$0.7661(\pm 0.0055)$	$0.7482(\pm 0.0104)$	$0.7520(\pm 0.0056)$	$0.7077(\pm 0.0213)$	**0.7653**(± 0.0086)



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

Figure 1: Visualizing Feature Matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ via T-SNE on ORL and Yale Datasets

2.4 Experimental Results on Noisy Data

To demonstrate the robustness of the proposed method, we also include the experimental results on image data corrupted by various noise. For each type we corrupt images successively to generate $\tilde{\mathbf{X}}$, $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ (which includes three stages of noise). In detail, four types of noise are considered: salt &

pepper noise, in which we corrupt 10% pixels in each stage and the results are summarized in Table 4 and Table 5; random corrupted pixels, in which we corrupt 20%, 5%, and 5% pixels in each stage and the results are summarized in Table 6 and Table 7;random corrupted regular patches, in which we corrupt 50%, 25%, and 25% pixels in each stage, and random corrupted irregular patches, in which we corrupt 20% pixels in each stage. We summarize the patch noise in Table 8, Table 9,Table 10 and Table **??**. The robustness of the proposed method is clearly verified according to these tables.

Table 4: ACC of Real Datasets with Salt & Pepper Noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	$0.7893(\pm0.0342)$	$0.8067(\pm0.0464)$	$0.8093(\pm0.0379)$	$0.8080(\pm0.0477)$	$0.8067(\pm0.0254)$	**0.8160**(±0.0421)
Yale	$0.3503(\pm0.0259)$	$0.3879(\pm0.0321)$	$0.3527(\pm0.0248)$	$0.3806(\pm0.0168)$	$0.3576(\pm0.0223)$	**0.4036**(±0.0180)
UMIST	$0.4734(\pm0.0157)$	$0.4800(\pm0.0150)$	$0.4602(\pm0.0268)$	$0.4814(\pm0.0086)$	$0.4275(\pm0.0162)$	**0.5078**(±0.0124)
ORL	$0.5720(\pm0.0195)$	$0.6155(\pm0.0252)$	$0.5475(\pm0.0083)$	$0.6225(\pm0.0275)$	$0.5350(\pm0.0173)$	**0.6670**(±0.0248)
COIL-20	$0.6678(\pm0.0123)$	$0.6723(\pm0.0247)$	$0.6547(\pm0.0190)$	$0.6762(\pm0.0175)$	$0.6782(\pm0.0316)$	**0.6830**(±0.0194)
USPS	$0.7706(\pm0.0008)$	$0.7542(\pm0.0004)$	$0.7716(\pm0.0003)$	$0.7592(\pm0.0003)$	$0.7505(\pm0.0058)$	**0.7793**(±0.0002)

Table 5: NMI of Real Datasets with Salt & Pepper Noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
Yale	$0.4008(\pm0.0268)$	$0.4393(\pm0.0261)$	$0.4070(\pm0.0277)$	$0.4255(\pm0.0229)$	$0.4144(\pm0.0203)$	**0.4509**(±0.0168)
UMIST	$0.5917(\pm0.0123)$	$0.5941(\pm0.0168)$	$0.5821(\pm0.0198)$	$0.5959(\pm0.0098)$	$0.5467(\pm0.0157)$	**0.6123**(±0.0083)
ORL	$0.7651(\pm0.0136)$	$0.7807(\pm0.0132)$	$0.7410(\pm0.0049)$	$0.7844(\pm0.0234)$	$0.7265(\pm0.0156)$	**0.8077**(±0.0154)
MNIST	$0.7464(\pm0.0175)$	$0.7648(\pm0.0241)$	$0.7689(\pm0.0217)$	$0.7676(\pm0.0242)$	$0.7580(\pm0.0179)$	**0.7776**(±0.0252)
USPS	$0.6565(\pm0.0011)$	$0.6349(\pm0.0003)$	$0.6581(\pm0.0002)$	$0.6478(\pm0.0001)$	$0.6356(\pm0.0053)$	**0.6594**(±0.0006)
COIL-20	$0.7528(\pm0.0149)$	$0.7506(\pm0.0174)$	$0.7497(\pm0.0131)$	$0.7546(\pm0.0079)$	$0.7512(\pm0.0136)$	**0.7584**(±0.0173)

Table 6: ACC of Real Datasets with Corrupt Pixel Noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
Yale	$0.3188(\pm0.0355)$	$0.3867(\pm0.0301)$	$0.3261(\pm0.0464)$	$0.3576(\pm0.0424)$	$0.3394(\pm0.0346)$	**0.4012**(±0.0286)
UMIST	$0.4557(\pm0.0134)$	$0.4706(\pm0.0261)$	$0.4483(\pm0.0170)$	$0.4720(\pm0.0235)$	$0.4310(\pm0.0269)$	**0.4866**(±0.0223)
ORL	$0.4775(\pm0.0137)$	$0.5370(\pm0.0141)$	$0.4850(\pm0.0157)$	$0.5145(\pm0.0192)$	$0.4650(\pm0.0190)$	**0.5600**(±0.0366)
MNIST	$0.7547(\pm0.0145)$	$0.7880(\pm0.0417)$	$0.7733(\pm0.0194)$	$0.7800(\pm0.0481)$	$0.7453(\pm0.0202)$	**0.8027**(±0.0293)
USPS	$0.7619(\pm0.0007)$	$0.7520(\pm0.0002)$	$0.7638(\pm0.0009)$	$0.7527(\pm0.0007)$	$0.7269(\pm0.0003)$	**0.7654**(±0.0006)
COIL-20	$0.6332(\pm0.0262)$	$0.6829(\pm0.0117)$	$0.6469(\pm0.0209)$	$0.6850(\pm0.0293)$	$0.6229(\pm0.0345)$	**0.6924**(±0.0337)

Figure 2 provides the visual decomposition results on Yale and ORL under different noise. For each figure, we present the noise-free images, $\mathbf{Z}$, $\tilde{\mathbf{Z}}$, $\mathbf{A}$ and $\tilde{\mathbf{A}}$ from top to bottom. This results indicate that our method can successfully avoid redundancy of features and learn the desire feature matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ with good representation performance.

2.5 Convergence Curves

We empirically validate the convergence analysis established in Section 5 using three datasets, ORL, Yale and BBCSports. We first learn the feature matrix $\mathbf{A}^k$ and weight matrix $\mathbf{X}^k$ for each iteration, then calculate the reconstruction error by $\Delta^k = \|\mathbf{Y} - \mathbf{A}^k\mathbf{X}^k\|_F^2$. The convergence curves of reconstruction error are plotted in Fig. 3. It can be found that our method converges to the optimal value around the tenth iteration on these datasets.

2.6 The Comparison with Deep Clustering Methods

In this subsection, we compare the proposed method with two classic deep clustering methods: NMF-D [27] and DEC [29]. We use regular and irregular patch noise settings as in the above experiments. The detailed results are summarized in Tables 12-15. It can be observed that our method performs better than these two compared methods.

References

[S1] Maurer A, Pontil M. K-dimensional coding schemes in Hilbert spaces[J]. IEEE Transactions on Information Theory, 2010, 56(11): 5839-5846.

[S2] Liu T, Gong M, Tao D. Large-cone nonnegative matrix factorization[J]. IEEE transactions on neural networks and learning systems, 2016, 28(9): 2129-2142.

Table 7: NMI of Real Datasets with Corrupt Pixel Noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
Yale	0.3618(±0.0334)	0.4200(±0.0274)	0.3755(±0.0466)	0.3983(±0.0360)	0.3908(±0.0287)	**0.4332**(±0.0226)
UMIST	0.5627(±0.0061)	0.5760(±0.0162)	0.5542(±0.0115)	0.5782(±0.0143)	0.5416(±0.0247)	**0.5857**(±0.0147)
ORL	0.6868(±0.0120)	0.7195(±0.0148)	0.6830(±0.0132)	0.7120(±0.0134)	0.6722(±0.0161)	**0.7295**(±0.0240)
MNIST	0.7170(±0.0278)	0.7576(±0.0306)	0.7540(±0.0202)	0.7534(±0.0364)	0.7192(±0.0180)	**0.7642**(±0.0144)
USPS	0.6487(±0.0007)	0.6358(±0.0002)	0.6486(±0.0008)	0.6358(±0.0005)	0.6079(±0.0007)	**0.6495**(±0.0011)
COIL-20	0.7548(±0.0139)	0.7670(±0.0100)	0.7529(±0.0151)	0.7659(±0.0174)	0.7369(±0.0181)	**0.7685**(±0.0093)

Table 8: ACC of Real Datasets with regular patch noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	0.7947(±0.0296)	0.8107(±0.0494)	0.8040(±0.0376)	0.8093(±0.0543)	0.7920(±0.0311)	**0.8160**(±0.0423)
Yale	0.3464(±0.0308)	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)	0.4449(±0.0272)	0.4710(±0.0259)	0.4223(±0.0248)	**0.4828**(±0.0251)
ORL	0.5430(±0.0151)	0.5465(±0.0243)	0.5145(±0.0288)	0.5520(±0.0198)	0.4695(±0.0368)	**0.5680**(±0.0207)
COIL-20	0.5188(±0.0095)	0.5274(±0.0209)	0.5145(±0.0121)	0.5278(±0.0054)	0.5293(±0.0284)	**0.5315**(±0.0198)
USPS	0.5218(±0.0057)	0.5210(±0.0005)	0.5296(±0.0074)	0.5195(±0.0018)	0.5306(±0.0078)	**0.5327**(±0.0065)

Table 9: NMI of Real Datasets with regular patch Noise. The best results are marked in bold.

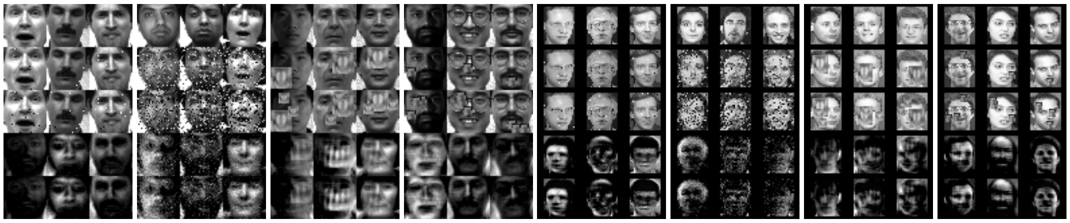
Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	0.7538(±0.0141)	0.7731(±0.0237)	0.7725(±0.0197)	0.7723(±0.0327)	0.7420(±0.0222)	**0.7775**(±0.0217)
Yale	0.3805(±0.0467)	0.3967(±0.0184)	0.3860(±0.0175)	0.3989(±0.0536)	0.3905(±0.0267)	**0.4166**(±0.0319)
UMIST	0.5311(±0.0247)	0.5458(±0.0167)	0.5204(±0.0260)	0.5424(±0.0207)	0.4942(±0.0210)	**0.5511**(±0.0209)
ORL	0.7117(±0.0150)	0.7146(±0.0169)	0.6876(±0.0266)	0.7159(±0.0140)	0.6579(±0.0334)	**0.7257**(±0.0141)
COIL-20	0.6104(±0.0147)	0.6073(±0.0153)	0.6138(±0.0119)	0.6027(±0.0064)	**0.6157**(±0.0218)	**0.6157**(±0.0155)
USPS	0.3924(±0.0079)	0.3720(±0.0031)	0.3846(±0.0039)	0.3730(±0.0065)	0.3703(±0.0173)	**0.4001**(±0.0137)

Table 10: ACC of Real Datasets with irregular patch Noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	0.2467(±0.0125)	0.2493(±0.0037)	0.2440(±0.0060)	0.2427(±0.0060)	0.2480(±0.0056)	**0.2497**(±0.0163)
Yale	0.4897(±0.0476)	0.5468(±0.0261)	0.4982(±0.0458)	0.5406(±0.0266)	0.4861(±0.0262)	**0.5549**(±0.0301)
UMIST	0.2174(±0.0044)	0.2247(±0.0056)	0.2115(±0.0089)	0.2235(±0.0087)	0.2136(±0.0051)	**0.2271**(±0.0057)
ORL	0.3155(±0.0076)	0.3250(±0.0127)	0.2880(±0.0132)	0.3230(±0.0110)	0.2780(±0.0082)	**0.3485**(±0.0146)
COIL-20	0.6747(±0.0195)	0.6792(±0.0202)	0.6706(±0.0228)	0.6782(±0.0181)	0.6586(±0.0177)	**0.6827**(±0.0145)
USPS	0.7548(±0.0005)	0.7388(±0.0002)	**0.7602**(±0.0001)	0.7455(±0.0001)	0.7320(±0.0001)	0.7559(±0.0003)

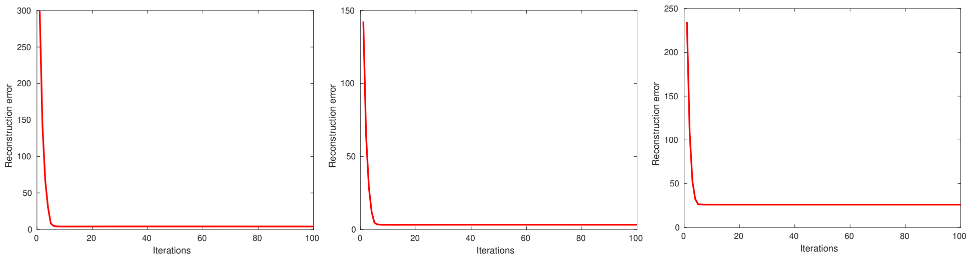
Table 11: NMI of Real Datasets with irregular patch Noise. The best results are marked in bold.

Dataset	K-Means	NMF	ONMF	$\ell_{2,1}$ NMF	CNMF	ANMF
MNIST	0.2788(±0.0095)	0.2768(±0.0086)	0.2811(±0.0101)	0.2772(±0.0118)	0.2665(±0.0104)	**0.2854**(±0.0147)
Yale	0.6078(±0.0248)	0.6543(±0.0199)	0.6119(±0.0260)	0.6462(±0.0215)	0.5931(±0.0124)	**0.6575**(±0.0205)
UMIST	0.3125(±0.0131)	0.3243(±0.0087)	0.2968(±0.0138)	0.3262(±0.0111)	0.2948(±0.0213)	**0.3359**(±0.0127)
ORL	0.5884(±0.0079)	0.6005(±0.0095)	0.5569(±0.0141)	0.6028(±0.0090)	0.5293(±0.0082)	**0.6230**(±0.0053)
COIL-20	0.7516(±0.0173)	0.7493(±0.0164)	0.7508(±0.096)	0.7484(±0.0124)	0.7340(±0.0132)	**0.7522**(±0.0137)
USPS	0.6409(±0.0006)	0.6234(±0.0002)	0.6395(±0.0001)	0.6286(±0.0002)	0.6189(±0.0003)	**0.6417**(±0.0004)



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

Figure 2: yale and ORL, row from top to bottom: origin, noisy data, noisy Z, A, A hat



(a) ORL (b) Yale (c) Yale

Figure 3: The Convergence of our Method on ORL and Yale Dataset

Table 12: ACC of Real Datasets with regular patch noise. The best results are marked in bold.

Dataset	NMF-D	DEC	ANMF
MNIST	0.7832(±0.0631)	0.7973(±0.0421)	**0.8160**(±0.0423)
Yale	0.3631(±0.0415)	0.3782(±0.0239)	**0.3852**(±0.0273)
UMIST	0.4693(±0.0131)	0.4754(±0.0662)	**0.4828**(±0.0251)
ORL	0.5483(±0.0131)	0.5187(±0.0241)	**0.5680**(±0.0207)
COIL-20	0.5233(±0.0164)	0.5167(±0.0146)	**0.5315**(±0.0198)
USPS	0.5225(±0.0022)	0.5320(±0.0043)	**0.5327**(±0.0065)

Table 13: NMI of Real Datasets with regular patch Noise. The best results are marked in bold.

Dataset	NMF-D	DEC	ANMF
MNIST	0.7631(±0.0282)	0.7725(±0.0452)	**0.7775**(±0.0217)
Yale	0.3836(±0.0725)	0.3998(±0.0153)	**0.4166**(±0.0319)
UMIST	0.5321(±0.0179)	0.4942(±0.0210)	**0.5511**(±0.0209)
ORL	0.7081(±0.0251)	0.6977(±0.0138)	**0.7257**(±0.0141)
COIL-20	0.5917(±0.0021)	0.6082(±0.0152)	**0.6157**(±0.0155)
USPS	0.3815(±0.0026)	0.3917(±0.0279)	**0.4001**(±0.0137)

Table 14: ACC of Real Datasets with irregular patch Noise. The best results are marked in bold.

Dataset	NMF-D	DEC	ANMF
MNIST	0.2372(±0.0045)	0.2464(±0.0051)	**0.2497**(±0.0163)
Yale	0.5379(±0.0386)	0.4926(±0.0543)	**0.5549**(±0.0301)
UMIST	0.2189(±0.0052)	0.2179(±0.0211)	**0.2271**(±0.0057)
ORL	0.3217(±0.0123)	0.2988(±0.0091)	**0.3485**(±0.0146)
COIL-20	0.6615(±0.0234)	0.6698(±0.0237)	**0.6827**(±0.0145)
USPS	0.7385(±0.0021)	0.7417(±0.0002)	**0.7559** (±0.0003)

Table 15: NMI of Real Datasets with irregular patch Noise. The best results are marked in bold.

Dataset	NMF-D	DEC	ANMF
MNIST	0.2631(±0.0232)	0.2679(±0.0112)	**0.2854**(±0.0147)
Yale	0.6422(±0.0198)	0.6278(±0.0297)	**0.6575**(±0.0205)
UMIST	0.3190(±0.0215)	0.3008(±0.0191)	**0.3359**(±0.0127)
ORL	0.5981(±0.0062)	0.6091(±0.0021)	**0.6230**(±0.0053)
COIL-20	0.7343(±0.0342)	0.7465(±0.0432)	**0.7522**(±0.0137)
USPS	0.6366(±0.0003)	0.6072(±0.0013)	**0.6417**(±0.0004)