Supplementary Material for A Deep Semi-NMF Model for Learning Hidden Representations

George Trigeorgis GEORGE.TRIGEORGIS08@IMPERIAL.AC.UK

Konstantinos Bousmalis K.BOUSMALIS@IMPERIAL.AC.UK

Stefanos Zafeiriou S.ZAFEIRIOU@IMPERIAL.AC.UK

Björn W. Schuller BJOERN.SCHULLER@IMPERIAL.AC.UK

Department of Computing, Imperial College London, United Kingdom

1. Analysis

Theorem 1. *(α) Fixing for any $\boldsymbol{Z}_i$ the residual*

$$\|\boldsymbol{X} - \boldsymbol{Z}_1 \cdots \boldsymbol{Z}_m \boldsymbol{H}_m\|_F^2 \tag{1}$$

decreases monotonically under the update rule for $\boldsymbol{H}$ and similarly, (β) fixing $\boldsymbol{H}$, the update rule for Z_i results in minimizing the objective function.

To prove (α), we now fix all the weights $\boldsymbol{Z}_i$ and solve for H imposing non-negativity constrains. To show that the final factors of the solution for the update rule for $\boldsymbol{H}$ (Equation 3) is correct we can show that at convergence it satisfies the KarushKuhnTucker (KKT) condition. Additionally, we show that the iteration of the update rule for H converges.

Proposition 1. *The limiting solution for the update rule for $\boldsymbol{H}$ satisfies the KKT condition.*

Proof. We introduce the Lagrangian function with the Lagrangian multipliers β_{ij}, in order to enforce the non-negativity constrains for matrix $\boldsymbol{H}$.

$$L(\boldsymbol{H}) = \text{tr}(-2\boldsymbol{X}^\top \prod_{k=1}^{i} \boldsymbol{Z}_k \boldsymbol{H}_i + 2\boldsymbol{H}^\top \prod_{k=i}^{1} \boldsymbol{Z}_k^\top \prod_{k=1}^{i} \boldsymbol{Z}_k - \beta \boldsymbol{H}^\top) \tag{2}$$

The fixed point equation that results from the gradient of the Lagrangian L is $\dfrac{\partial \boldsymbol{L}}{\partial \boldsymbol{H}} = -2\Psi^\top \boldsymbol{X} + 2\boldsymbol{H}^\top \Psi\Psi^\top - \beta = 0$. From the complementary slackness condition, we obtain $(-2\boldsymbol{X}\boldsymbol{\Psi}^\top + 2\boldsymbol{H}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top)_{ik}\boldsymbol{H}_{ik} = \beta_{ik}\boldsymbol{H}_{ik} = 0$. At convergence we have $\boldsymbol{H}^{(\text{inf})} = \boldsymbol{H}^{(t+1)} = \boldsymbol{H}^{(t)} = \boldsymbol{H}$, as the update rule satisfies the fixed point equation conditions.

$$\boldsymbol{H}_i = \boldsymbol{H}_i \odot \sqrt{\frac{[\boldsymbol{\Psi}^\top \boldsymbol{X}]^{\text{pos}} + [\boldsymbol{\Psi}^\top \boldsymbol{\Psi}]^{\text{neg}} \boldsymbol{H}_i}{[\boldsymbol{\Psi}^\top \boldsymbol{X}]^{\text{neg}} + [\boldsymbol{\Psi}^\top \boldsymbol{\Psi}]^{\text{pos}} \boldsymbol{H}_i}}. \tag{3}$$

As $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = [\boldsymbol{\Psi}\boldsymbol{\Psi}^\top]^{\text{pos}} - [\boldsymbol{\Psi}\boldsymbol{\Psi}^\top]^{\text{neg}}$ and similarly $\boldsymbol{\Psi}^\top \boldsymbol{X} = [\boldsymbol{\Psi}^\top \boldsymbol{X}]^{\text{pos}} - [\boldsymbol{\Psi}^\top \boldsymbol{X}]^{\text{neg}}$, then the Eq. 3 reduces to

$$(-2\boldsymbol{X}\boldsymbol{\Psi}^\top + 2\boldsymbol{H}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top)_{ik}\boldsymbol{H}_{ik}^2 = 0. \tag{4}$$

Thus, Equation 4 is equivalent to Equation 3 as the first factor is similar and the second factor for Equation 4 is zero exactly when Equation 3 is zero and vice versa. ■

Proposition 2. *The residual Equation 1 is monotonically decreasing under the update rule of Equation 3 for Zi for any layer i.*

Proof. We rewrite the objective function J as in (Lee & Seung, 2001)

$$G = Tr[-2\boldsymbol{B}^{\top(\text{pos})}\boldsymbol{G} + 2\boldsymbol{B}^{\top(\text{neg})}\boldsymbol{G} + \boldsymbol{G}\boldsymbol{A}^{\top(\text{pos})}\boldsymbol{G}^{\top} + \boldsymbol{G}\boldsymbol{A}^{\top(\text{neg})}\boldsymbol{G}^{\top}] \tag{5}$$

where $\boldsymbol{A} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}, \boldsymbol{B} = \boldsymbol{X}^{\top}\boldsymbol{\Psi}$ and $\boldsymbol{G} = \boldsymbol{H}$.

Using auxiliary functions F and G, as in (Lee & Seung, 2001) we define:

Definition 1.1. *G(h, h') is an auxiliary function for F(h) if the conditions*

$$G(h, h') \geq F(h), \quad G(h, h) = F(h) \tag{6}$$

are satisfied.

Lemma 1.2. *If G is an auxiliary function, the F is non-increasing under the the update*

$$h^{t+1} = \arg\min_{h} G(h, h^t) \tag{7}$$

Proof. $F(h^{(t+1)} \leq G(h^{(t+1)}, h^t) \leq G(h^t, h^t) = F(h^t)$ ■

Thus G is a monotonically decreasing function if we find an appropriate auxiliary function satisfying the conditions. Using the auxiliary function

$$G(\boldsymbol{G}, \boldsymbol{G}') = -\sum_{ik} 2\boldsymbol{B}^{\text{pos}}_{ik}\boldsymbol{G}'_{ik}(1 + \log\frac{\boldsymbol{G}_{ik}}{\boldsymbol{G}'_{ik}}) + \sum_{ik}\boldsymbol{B}^{\text{neg}}_{ik}\frac{\boldsymbol{G}^2_{ik} + \boldsymbol{G}'^2_{ik}}{\boldsymbol{G}'_{ik}} \tag{8}$$

$$+ \sum_{ik}\frac{(\boldsymbol{A}^{\text{pos}}_{ik}\boldsymbol{H}')_{ik}\boldsymbol{H}^2_{ik}}{\boldsymbol{H}'_{ik}} - \sum_{ikl}\boldsymbol{A}^{\text{neg}}_{kl}\boldsymbol{H}'_{il}\boldsymbol{H}'_{il}(1 + \log\frac{\boldsymbol{H}_{ik}\boldsymbol{H}_{il}}{\boldsymbol{H}'_{ik}\boldsymbol{H}'_{il}}) \tag{9}$$

as in (Ding et al., 2010) we can prove that G is such a function. $F(\boldsymbol{G}, \boldsymbol{G}')$ satisfies the necessary requirements and furthermore it's a convex function in $\boldsymbol{G}$ and its global minima is

$$\boldsymbol{G} = \arg\min_{h} G(h, h^t) = \boldsymbol{G}_i \odot \sqrt{\frac{[\boldsymbol{\Psi}^{\top}\boldsymbol{X}]^{\text{neg}} + [\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}]^{\text{neg}}\boldsymbol{G}_i}{[\boldsymbol{\Psi}^{\top}\boldsymbol{X}]^{\text{neg}} + [\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}]^{\text{pos}}\boldsymbol{G}_i}}. \tag{10}$$

as been proved by (Ding et al., 2010).

■

We prove (β) by fixing the rest of the weights for the i_{th} layer using the fact that $\boldsymbol{Z}$ is minimal when $\partial C/\partial \boldsymbol{Z}_i = 0$. That is: $-\boldsymbol{Z}^{\top}_{i-1}\cdots\boldsymbol{Z}^{\top}_1\boldsymbol{X}\tilde{\boldsymbol{H}}^{\top}_i + \boldsymbol{Z}^{\top}_{i-1}\boldsymbol{Z}^{\top}_{i-2}\cdots\boldsymbol{Z}^{\top}_1\boldsymbol{Z}_1\cdots\boldsymbol{Z}_{i-1}\boldsymbol{Z}_i\tilde{\boldsymbol{H}}_i\tilde{\boldsymbol{H}}_i{}^{\top} = 0$.

$$\begin{aligned}\boldsymbol{Z}_i &= (\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^{\top}X\tilde{\boldsymbol{H}}_i{}^{\top}(\tilde{\boldsymbol{H}}_i\tilde{\boldsymbol{H}}_i{}^{\top})^{-1} \\ \boldsymbol{Z}_i &= \boldsymbol{\Psi}^{\dagger}\boldsymbol{X}\tilde{\boldsymbol{H}}^{\dagger}_i\end{aligned} \tag{11}$$

where $\boldsymbol{\Psi} = \boldsymbol{Z}_1\cdots\boldsymbol{Z}_{i-1}$.

In this Supplementary Material, we provide experimental results in regards to the reconstruction error of each of the tested NMF solvers, their variants, Multi-Layer NMF and Deep Semi-NMF. We show the mean reconstruction error for each of the algorithms with a variable number of components. The error quantifies the deviation of the reconstruction with the original matrix $\boldsymbol{X}$, using the Euclidean objective function: $\frac{1}{N}\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|^2$, with N amount of samples. In all of the experiments we see that Deep Semi-NMF has a comparable reconstruction error to that of Semi-NMF, which is in contrast with the multi–layer NMF and GNMF which sacrifice the reconstruction quality, in return for uncovering more meaningful features.

	Number of components										
Name	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**
CMU MultiPIE—Pixel Intensities											
NMF (MUL)	2.91	2.80	2.68	2.59	2.52	2.45	2.41	2.36	2.32	2.28	2.23
NeNMF	2.69	2.53	2.40	2.29	2.20	2.12	2.04	1.98	1.92	1.86	1.80
GNMF	3.16	3.10	3.05	3.04	3.03	3.01	2.98	3.00	3.02	3.02	3.02
Semi-NMF	2.66	2.50	2.37	2.26	2.16	2.08	2.01	1.95	1.89	1.83	1.78
Multi-layer NMF	2.96	2.83	2.71	2.62	2.54	2.47	2.42	2.36	2.31	2.26	2.22
Deep Semi-NMF	2.69	2.53	2.40	2.30	2.20	2.13	2.07	2.01	1.96	1.90	1.86
CMU MultiPIE—Image Gradient Orientations											
Semi-NMF	0.14	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.12	0.12	0.12
Deep Semi-NMF	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.12	0.12
CMU PIE Pose dataset											
NMF (MUL)	10.53	9.84	9.36	8.85	8.51	8.18	7.91	7.64	7.42	7.20	7.00
NeNMF	9.83	9.06	8.39	7.87	7.39	6.97	6.60	6.24	5.94	5.61	5.36
GNMF	10.56	9.96	9.35	9.04	8.73	8.46	8.18	7.99	7.81	7.64	7.48
Semi-NMF	9.14	8.28	7.57	6.95	6.43	5.96	5.53	5.12	4.76	4.42	4.13
Multi-layer NMF	11.11	10.58	10.16	9.56	9.28	8.98	8.49	7.97	7.63	7.30	6.98
Deep Semi-NMF	9.18	8.31	7.61	7.01	6.50	6.02	5.67	5.28	4.99	4.70	4.39
CMU PIE Pose using IGO features dataset											
Deep Semi-NMF	0.35	0.33	0.32	0.31	0.29	0.28	0.27	0.27	0.26	0.25	0.24
Semi-NMF	0.35	0.33	0.32	0.30	0.29	0.28	0.27	0.26	0.25	0.24	0.24
XM2VTS dataset											
NMF (MUL)	9.20	8.82	8.51	8.29	8.08	7.91	7.75	7.61	7.49	7.38	7.28
NeNMF	8.45	7.99	7.59	7.27	6.98	6.74	6.52	6.31	6.12	5.95	5.79
GNMF	10.60	10.48	10.36	10.33	10.31	10.28	10.25	10.26	10.27	10.32	10.36
Semi-NMF	8.43	7.95	7.55	7.22	6.93	6.68	6.46	6.25	6.05	5.88	5.72
Multi-layer NMF	9.17	8.86	8.52	8.19	7.96	7.79	7.59	7.44	7.32	7.20	7.20
Deep Semi-NMF	8.48	8.01	7.61	7.29	7.00	6.75	6.52	6.32	6.14	5.98	5.82
XM2VTS using IGO features dataset											
Semi-NMF	0.46	0.45	0.44	0.43	0.42	0.42	0.41	0.40	0.40	0.39	0.39
Deep Semi-NMF	0.46	0.45	0.44	0.43	0.42	0.42	0.41	0.40	0.40	0.39	0.39

References

Ding, Chris HQ, Li, Tao, and Jordan, Michael I. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010.

Lee, Daniel D. and Seung, H. Sebastian. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.