An Alternating Modulus Nonnegative Least Squares Method for Nonnegative Matrix Factorization

Ning Zheng[2] **Ken Hayami**[1,2] **Nobutaka Ono**[1,2]

1. National Institute of Informatics, Japan
2. SOKENDAI (The Graduate University for Advanced Studies), Japan

Numerical Linear Algebra and Applications (NL2A)
CIRM Luminy, France
October 24 - 28, 2016

Outline

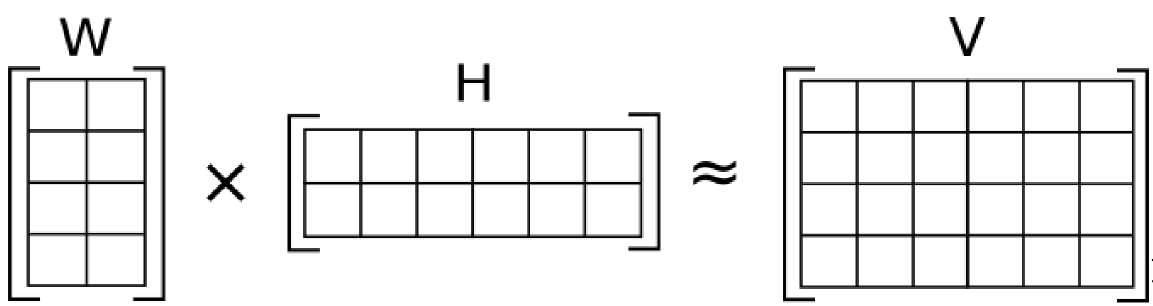
1. Problem
2. Alternating Nonnegative Least Squares Method
   - Multiplicative update method
   - Projection Gradient method
3. Alternating Modulus Nonnegative Least Squares Method
4. Numerical Results and Conclusion

Consider the Nonnegative Matrix Factorization (NMF)

$$\min f(W, H) := \frac{1}{2}\|V - WH\|_F^2,$$

- $V \in \mathbf{R}^{m \times n}$ is a given nonnegative matrix;
- $W \in \mathbf{R}^{m \times r}$ and $H \in \mathbf{R}^{r \times n}$ are unknown nonnegative matrices;
- Frobenius norm $\|A\|_F^2 = \sum_{i,j} a_{ij}^2$;
- $r \ll \min(m, n)$.

The NMF seeks a low rank approximation of a given nonnegative matrix.



[1]

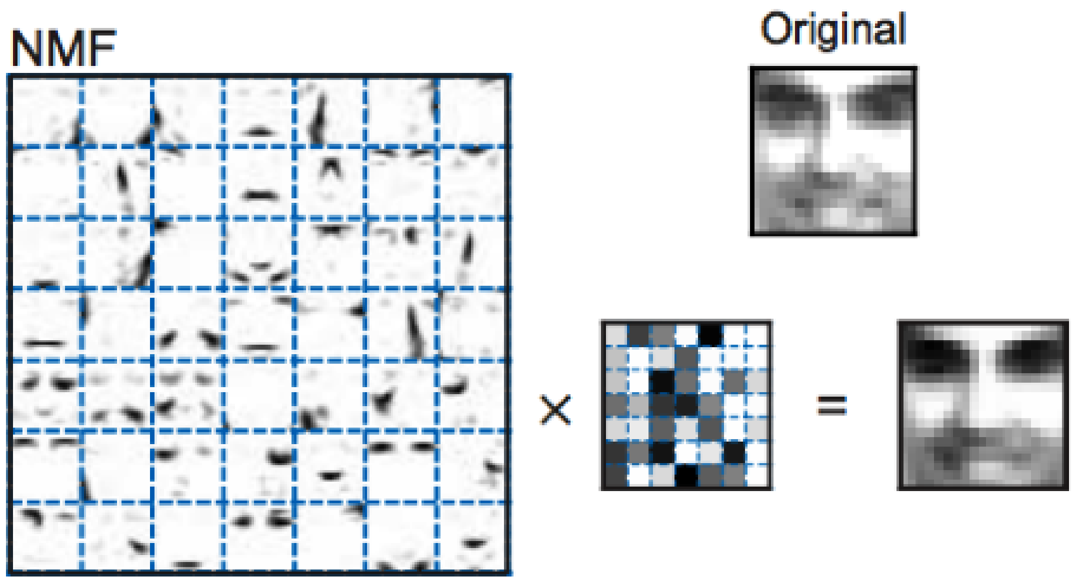
[1] https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

NMF problems arise in many scientific computing and engineering applications, e.g.,

- Computer vision,
- Spectral data analysis,
- Text mining,
- Document clustering,
- Chemometrics,
- Audio signal separation,
- Recommender systems,
- Image classification, etc.

Image Classification: Face Recognition

Extract features or individual components like nose, eyes and mouth from a face: (Lee and Seung, Nature, 99')



• NMF is non-convex. Let (W^*, H^*) be a pair of local minimizer or stationary point and $D \geq 0$ is a nonsingular matrix

$$f(W^*, H^*) = \frac{1}{2}\|V - W^*H^*\|_F^2 = \frac{1}{2}\|V - W^*DD^{-1}H^*\|_F^2 = f(W^*D, D^{-1}H^*).$$

• Alternating nonnegative least squares (ANLS) algorithm or two-block coordinate descent method

---

0. For $k = 0, 1, 2, \ldots$ until convergence
1. $\quad H^{k+1} = \mathrm{argmin}\|V - W^kH\|_F^2$ subject to $H \geq 0$
2. $\quad W^{k+1} = \mathrm{argmin}\|V - WH^{k+1}\|_F^2$ subject to $W \geq 0$

• The subproblems 1 and 2 are nonnegative constrained least squares (NNLS) problems, which are convex.

• (Grippo and Siandrone, 00') Any limit point of the sequence generated by the optimal solutions of each of the two subproblems is a stationary point of NMF.

• Gradient descent method for the subproblems: the objective function f decreases if one goes from x in the direction of the negative gradient of f at x.

• The gradient of $f(W, H)$ with respect to W and H are

$$\nabla_H f(W, H) = W^\mathsf{T}(WH - V)$$
$$\nabla_W f(W, H) = (WH - V)H^\mathsf{T}.$$

• Merit: monotonic decrease

$$f(W^{k+1}, H^k) \leq f(W^k, H^k) \quad \text{and} \quad f(W^{k+1}, H^{k+1}) \leq f(W^{k+1}, H^k)$$

• Demerit: the gradient descent method may suffer from the zigzag phenomenon when approaching the local minimizer if the condition number is bad.

- Gradient descent method 1: multiplicative update (MU) (Lee and Seung, 01').
  - ▷ It can be derived from the element-wise update

$$H_{ij} = H_{ij} - \eta_{ij}[\nabla_H f(W,H)]_{ij} = H_{ij} - \eta_{ij}[W^\top W H]_{ij} + \eta_{ij}[W^\top V]_{ij},$$
$$W_{ij} = W_{ij} - \xi_{ij}[\nabla_W f(W,H)]_{ij} = W_{ij} - \xi_{ij}[W H H^\top]_{ij} + \xi_{ij}[V H^\top]_{ij},$$

  - ▷ Zero the potentially negative part $H_{ij} - \eta_{ij}[W^\top W H]_{ij} = 0$, $W_{ij} - \xi_{ij}[W H H^\top]_{ij} = 0$,

$$\eta_{ij} = \frac{H_{ij}}{[W^\top W H]_{ij}} \quad \text{and} \quad \xi_{ij} = \frac{W_{ij}}{[W H H^\top]_{ij}}$$

  - ▷ We have

$$H_{ij} = \frac{H_{ij}[W^\top V]_{ij}}{[W^\top W H]_{ij}} \quad \text{and} \quad W_{ij} = \frac{W_{ij}[V H^\top]_{ij}}{[W H H^\top]_{ij}}.$$

• Gradient descent method 2: projected gradient (PG) (Bertsekas, 76').

▷ It can be derived from

$$H^{k+1} = P(H^k - \eta[\nabla_H f(W, H^k)]) \quad \text{and} \quad W^{k+1} = P(W^k - \xi[\nabla_H f(W^k, H)])$$

▷ The orthogonal projection operator $P(X)$ is the matrix whose (i, j)th component is the maximum of X_{ij} and 0.

▷ The choice of step size η and ξ are based on the Armijo condition or sufficient decrease condition on each column of H^k and each row of W^k, respectively.

▷ Take jth column of H^{k+1} for example, set $0 < \beta < 1$, $0 \leq \mu < 1$ and

$$h_j^{k+1} = P(h_j^k - \beta^m \eta_j^* [\nabla_H f(W, H^k)]_j),$$

and find the smallest integer $m \geq 0$ that satisfies the Armijo condition

$$\|v_j - W^k h_j^{k+1}\|_2^2 \leq \|v_j - W^k h_j^k\|_2^2 + 2\mu [\nabla_H f(W, H^k)]_j^\mathsf{T} (h_j^{k+1} - h_j^k).$$

• Other iterative methods for the subproblem NNLS

$$H^{k+1} = \text{argmin}\|V - W^k H\|_F^2 \quad \text{subject to} \quad H \geq 0$$
$$W^{k+1} = \text{argmin}\|V - W H^{k+1}\|_F^2 \quad \text{subject to} \quad W \geq 0$$

can be applied for NMF:

▷ Active set gradient descent (Lawson and Hanson, 74'; Kim and Park, 08');
▷ Block principal pivoting method (Kim and Park, 11');
▷ A new active set method (Hager and Zhang, 06'; Zhang, etc., 14');

• Possible new strategy

▷ Gradient projection conjugate gradient (GPCG) (Moré and Toraldo, 89')

• New strategy

▷ Modulus-type inner outer iteration method
▷ Hybrid modulus active set method (Zheng, Hayami and Yin, SIMAX, 16')

New Strategy

Nonnegativity: $h_j \geq 0$

$\downarrow$

Variable transformation

$h_j = g(\boldsymbol{z}_j)$,

$\downarrow$

Apply iterative methods on $\boldsymbol{z}_j$
to obtain an unconstrained solution sequence $\{\boldsymbol{z}_j^k\}_{k=0}^{+\infty}$

$\downarrow$

Update $h_j^k = g(\boldsymbol{z}_j^k)$
to obtain nonnegative constrained solution sequence $\{h_j^k\}_{k=0}^{+\infty}$

- Reflective Newton method (Coleman and Li, 96');
- Nonnegativity enforcement: $g(\boldsymbol{z}) = e^{\boldsymbol{z}}$ (Hanke and Nagy, 00');
- Modulus: $g(\boldsymbol{z}) = \boldsymbol{z}_j + |\boldsymbol{z}_j|$ (Van Bokhoven; Bai,10).

• Consider the solution of NNLS problem

$$H^{k+1} = \operatorname{argmin} \|V - W^k H\|_F^2 \quad \text{subject to} \quad H \geq 0.$$

• Set $H = [h_1, h_2, ..., h_n]$ and $V = [v_1, v_2, ..., v_n]$. If each column of H is updated independently, we only need to consider

$$\min \|v_j - W^k h_j\|_2^2 \quad \text{subject to} \quad h_j \geq 0,$$

where $j = 1, 2, ..., n$.

• Kuhn-Kurush-Tucker (KKT) conditions

$$h_j \geq 0, \quad [\nabla_H f(W^k, H)]_j = (W^k)^{\mathsf{T}}(W^k h_j - v_j) \geq 0 \quad \text{and} \quad h_j^{\mathsf{T}}[\nabla_H f(W^k, H)]_j = 0.$$

• Modulus-type inner outer iteration:

For $j = 1, 2, ..., n$, set $h_j = z_j + |z_j|$ and $[\nabla_H f(W^k, H)]_j = \Omega(|z_j| - z_j)$, the KKT conditions are equivalent to an implicit fixed-point equation

$$(\Omega + (W^k)^\mathsf{T} W^k) z_j = (\Omega - (W^k)^\mathsf{T} W^k)|z_j| + (W^k)^\mathsf{T} v_j.$$

Note that the fixed point iteration

$$(\Omega + (W^k)^\mathsf{T} W^k) z_j^{i+1} = (\Omega - (W^k)^\mathsf{T} W^k)|z_j^i| + (W^k)^\mathsf{T} v_j$$

is the normal equation of the unconstrained least squares problem

$$\min \left\| \begin{bmatrix} W^k \\ \Omega^{1/2} \end{bmatrix} z_j^{i+1} - \begin{bmatrix} -W^k |z_j^i| + v_j \\ \Omega^{1/2} |z_j^i| \end{bmatrix} \right\|_2 .$$

Set $Z = [z_1, z_2, ..., z_n]$, we have the following modulus-type inner outer iteration method for

$$\min \|V - W^k H\|_F^2 \quad \text{subject to} \quad H \geq 0.$$

---

0. For $i = 0, 1, 2, \ldots$ until convergence
1. Solve Z^{i+1} from

$$(\Omega + (W^k)^\mathsf{T} W^k) Z^{i+1} = (\Omega - (W^k)^\mathsf{T} W^k)|Z^i| + (W^k)^\mathsf{T} V,$$

or

$$\min \left\| \begin{bmatrix} W^k \\ \Omega^{1/2} \end{bmatrix} Z^{i+1} - \begin{bmatrix} -W^k |Z^i| + V \\ \Omega^{1/2} |Z^i| \end{bmatrix} \right\|_2.$$

2. Compute $H^{i+1} = Z^{i+1} + |Z^{i+1}|$.

CG for Inner Matrix System

- The solution of the normal matrix equation is required.
- We first review that for the solution of normal equation

$$\min \|Ax - b\|_2 \quad \Longleftrightarrow \quad A^\top A x = A^\top b,$$

the CGLS method (Hestenes, Stiefel, 52') is proposed as follows.

---

0. Choose initial x^0, $r^0 = b - Ax^0$, $s^0 = A^\top r^0$ and $p^0 = s^0$.
1. For $k = 0, 1, 2, \ldots$ until convergence
2. $\quad \alpha_k = (s^k, s^k)/(Ap^k, Ap^k)$
3. $\quad x^{k+1} = x^k + \alpha_k p^k$
4. $\quad r^{k+1} = r^k - \alpha_k A p^k$
5. $\quad s^{k+1} = A^\top r^{k+1}$
6. $\quad \beta_{k+1} = (s^{k+1}, s^{k+1})/(s^k, s^k)$
7. $\quad p^{k+1} = s^{k+1} + \beta_{k+1} p^k$

CG for Inner Matrix System

- Now we consider the solution of normal equation with multiple right hand sides

$$\min \|AX - B\|_F \quad \Longleftrightarrow \quad A^\mathsf{T}AX = A^\mathsf{T}B,$$

the CGLS method can be derived as follows.

---

0. Choose initial X^0, $R^0 = B - AX^0$, $S^0 = A^\mathsf{T}R^0$ and $P^0 = S^0$.
1. For $k = 0, 1, 2, \ldots$ until convergence
2. $\quad \Gamma_k = \mathrm{diag}((S^k)^\mathsf{T}(S^k))./\mathrm{diag}((AP^k)^\mathsf{T}(AP^k))$
3. $\quad X^{k+1} = X^k + P^k\Gamma_k$
4. $\quad R^{k+1} = R^k - AP^k\Gamma_k$
5. $\quad S^{k+1} = A^\mathsf{T}R^{k+1}$
6. $\quad \Lambda_{k+1} = \mathrm{diag}((S^{k+1})^\mathsf{T}(S^{k+1}))./\mathrm{diag}((S^k)^\mathsf{T}(S^k))$
7. $\quad P^{k+1} = S^{k+1} + P^k\Lambda_{k+1}$

- Convergence theorem

If W^k is full column rank, modulus-type inner outer iteration algorithm converges when the inner system is solved exactly, or iteratively with

$$\|e^k\|_\Omega \le \gamma^k \|\varepsilon^k\|_\Omega \quad \text{and} \quad \gamma^k < \frac{\alpha(1-\delta)}{\tau + c} \quad \text{for } k \ge k_0$$

where e^k and ε^k are stopping criteria of inner iteration and outer iteration, respectively, and k_0 is an integer, $0 \le \alpha < 1$,

$$\begin{aligned}
\tau &= \|(\Omega + (W^k)^\mathsf{T} W^k)^{-1}\|_{\Omega^{1/2},2} \|\Omega + (W^k)^\mathsf{T} W^k\|_{\Omega^{1/2},2} \\
\delta &= \|(\Omega + (W^k)^\mathsf{T} W^k)^{-1}(\Omega - (W^k)^\mathsf{T} W^k)\|_{\Omega^{1/2},2} \\
c &= \|(\Omega + (W^k)^\mathsf{T} W^k)^{-1}\|_{\Omega^{1/2},2} \|(\Omega - (W^k)^\mathsf{T} W^k)\|_{\Omega^{1/2},2}.
\end{aligned}$$

- Alternating modulus least squares (AMLS) method for NMF

---

0. For $k = 0, 1, 2, \dots$ until convergence
1. Solve $\min\|V - W^k H\|_F^2$ subject to $H \geq 0$ using modulus method
2. Solve $\min\|V - WH^{k+1}\|_F^2$ subject to $W \geq 0$ using modulus method

---

- Merit:

  ▷ Easy to implement

  ▷ Transform the nonnegative constrained least squares problem to a series of unconstrained least squares problems, which can be solved efficiently by CGLS, LSQR, BA-GMRES, etc (Morikuni and Hayami, 13').

- Demerit: the convergence rate of the fixed-point iteration is at best linear.

Numerical Experiments

• Compare the proposed modulus (Mod) method with the existing methods including multiplicative update (MU) method, projected gradient (PG) method, projected gradient method with Armijo condition (PGA).
• The testing problems contain

Synthetic data: Consider matrix V is randomly generated by the normal distribution with mean 0 and standard deviation 1

$$V_{ij} = |N(0,1)|.$$

The initial matrices are also constructed randomly. The size of the problem is $(m, r, n) = (100, 20, 500)$.

Image data: ORL face image database. $(m, r, n) = (10304, 25, 400)$.

Numerical Experiments

- MATLAB 7.8 with machine precision $\epsilon = 1.1 \times 10^{-16}$.
- The initial matrices were chosen to be random matrices. For the modulus-type iteration methods, the parameter matrix was chosen to be $\Omega = \omega I$, where ω is a positive parameter.
- The stopping criterion for the outer iteration of all methods is chosen as

$$\frac{|f(W^{k+1}, H^{k+1}) - f(W^k, H^k)|}{f(W^0, H^0)} < tol = 10^{-8}.$$

- In order to perform a fair comparison among different methods, the parameters are chosen as

$$\mu = 0.1, \quad \beta = 0.9 \quad \text{and} \quad \omega = 1$$

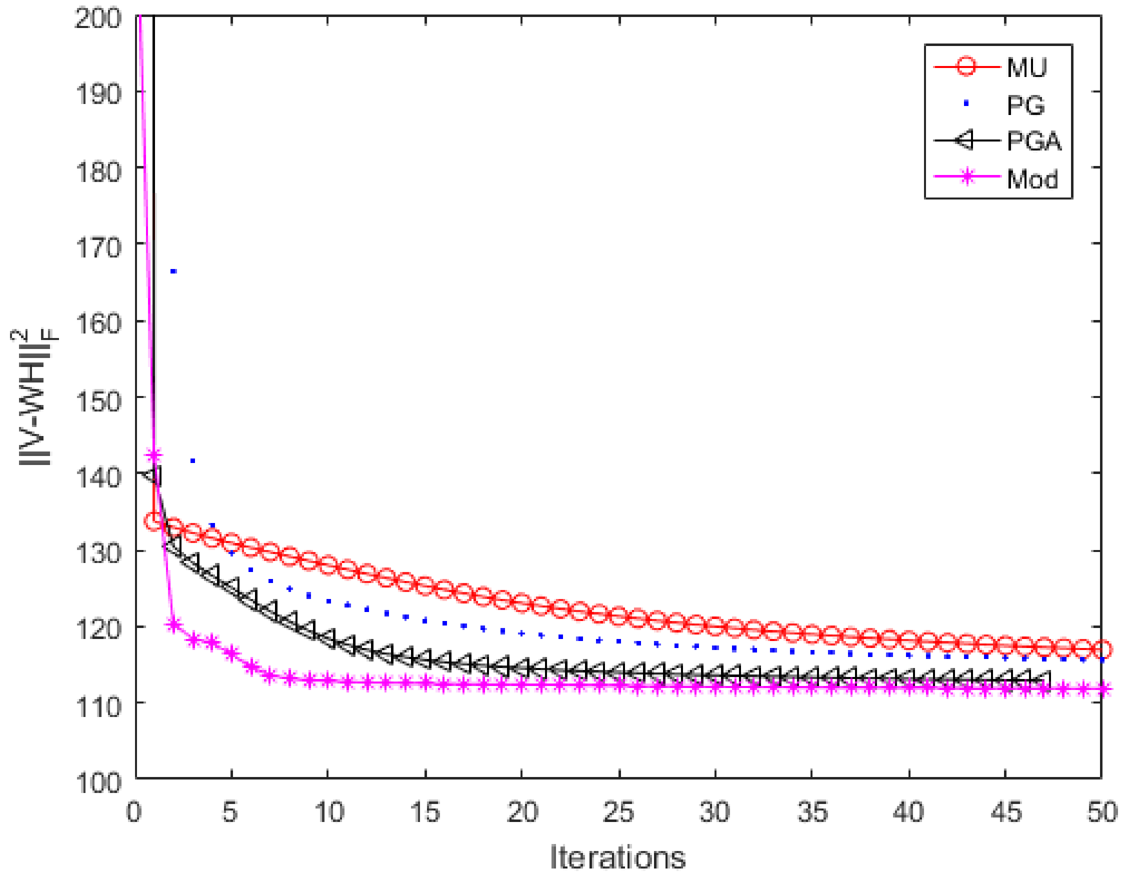
- The maximum number of iteration steps is restricted to be 5,000.

Synthetic data

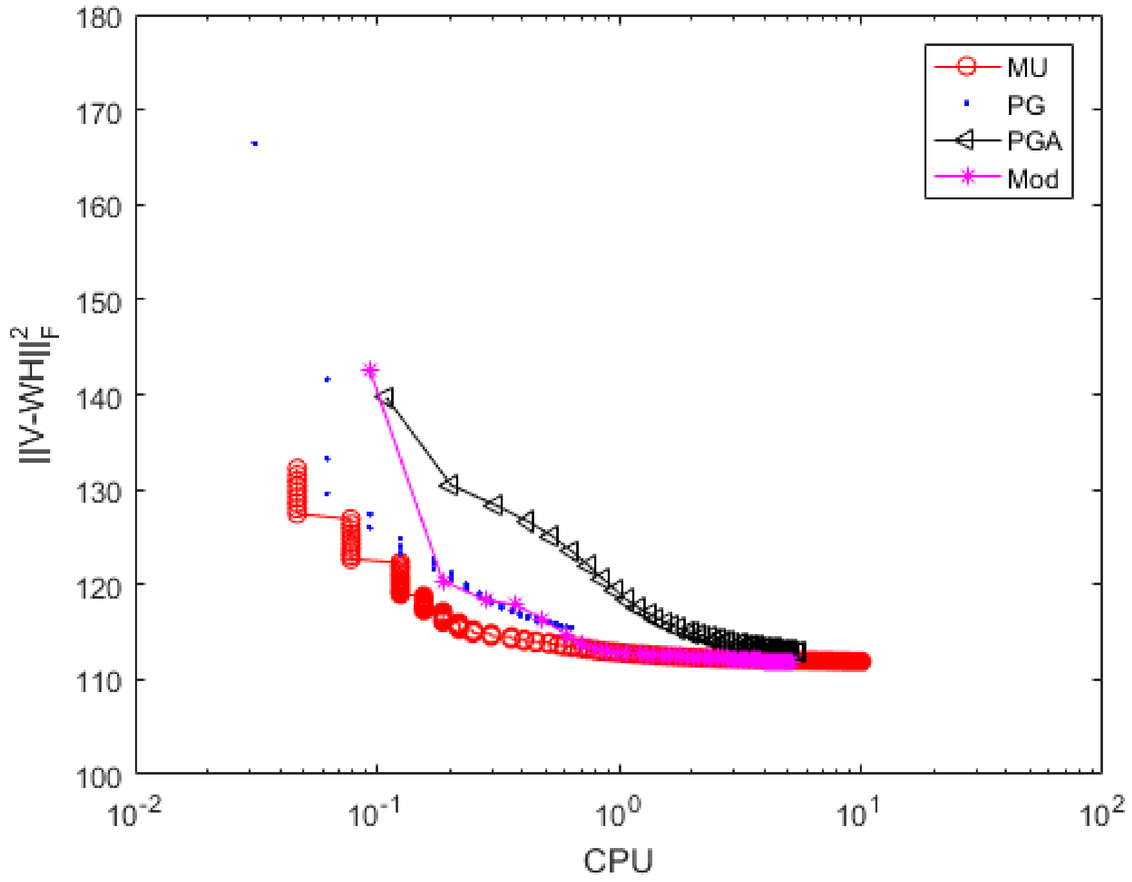
Comparison of the iterative methods for random problem.

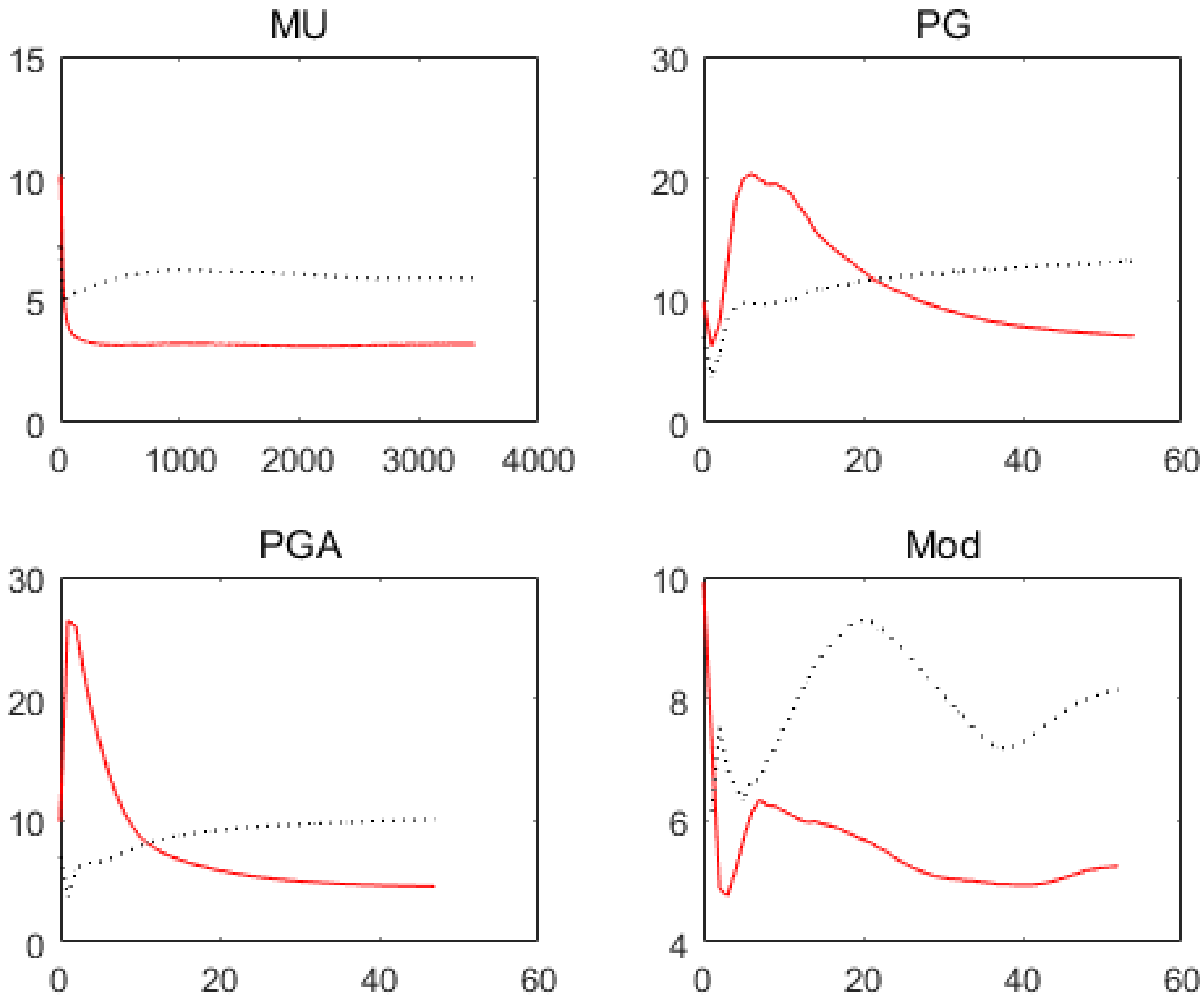
Methods	Iterations	$f(W,H)$	CPU
MU	3468	111.90	10.09
PG	54	115.47	0.64
PGA	47	112.93	5.52
Mod	52	111.88	5.14

Synthetic data: iterations



Synthetic data: computational time





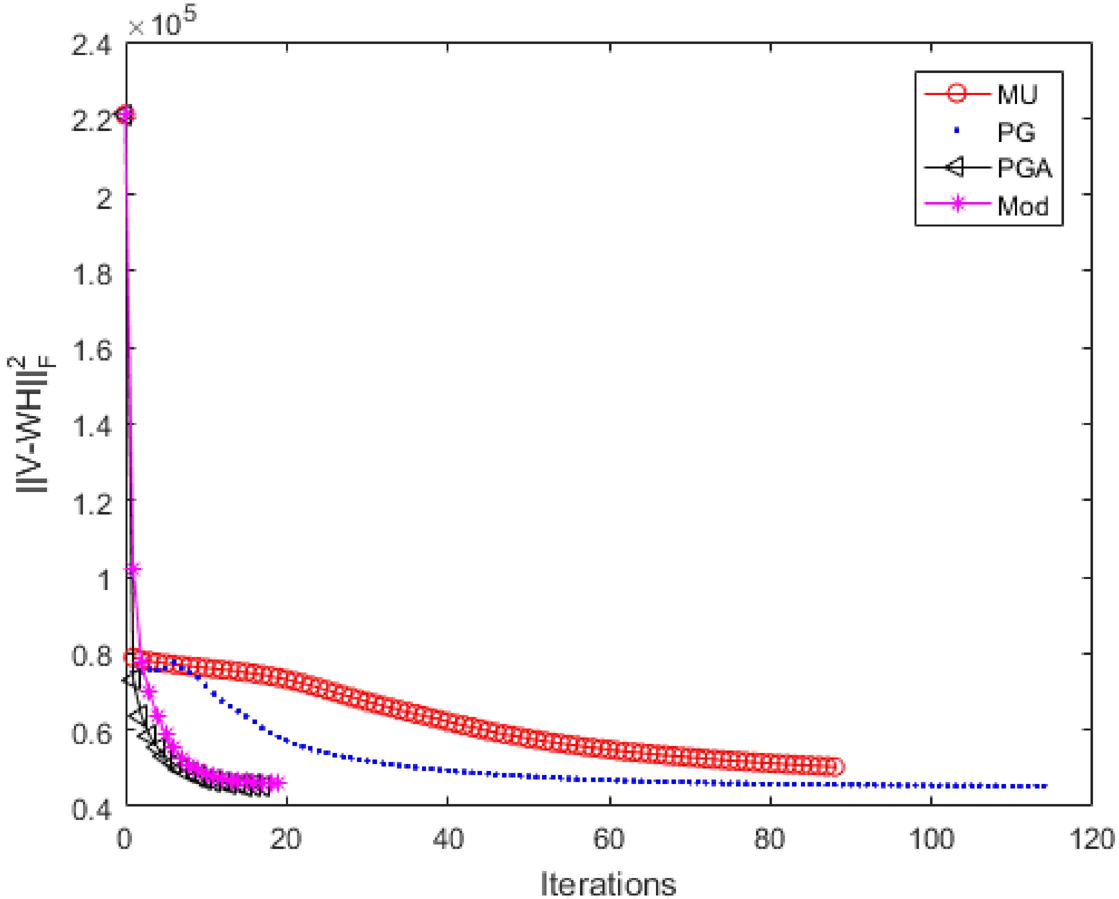
Iterations vs. condition numbers of W^k (red line) and H^k (black dot).

ORL facedata problem

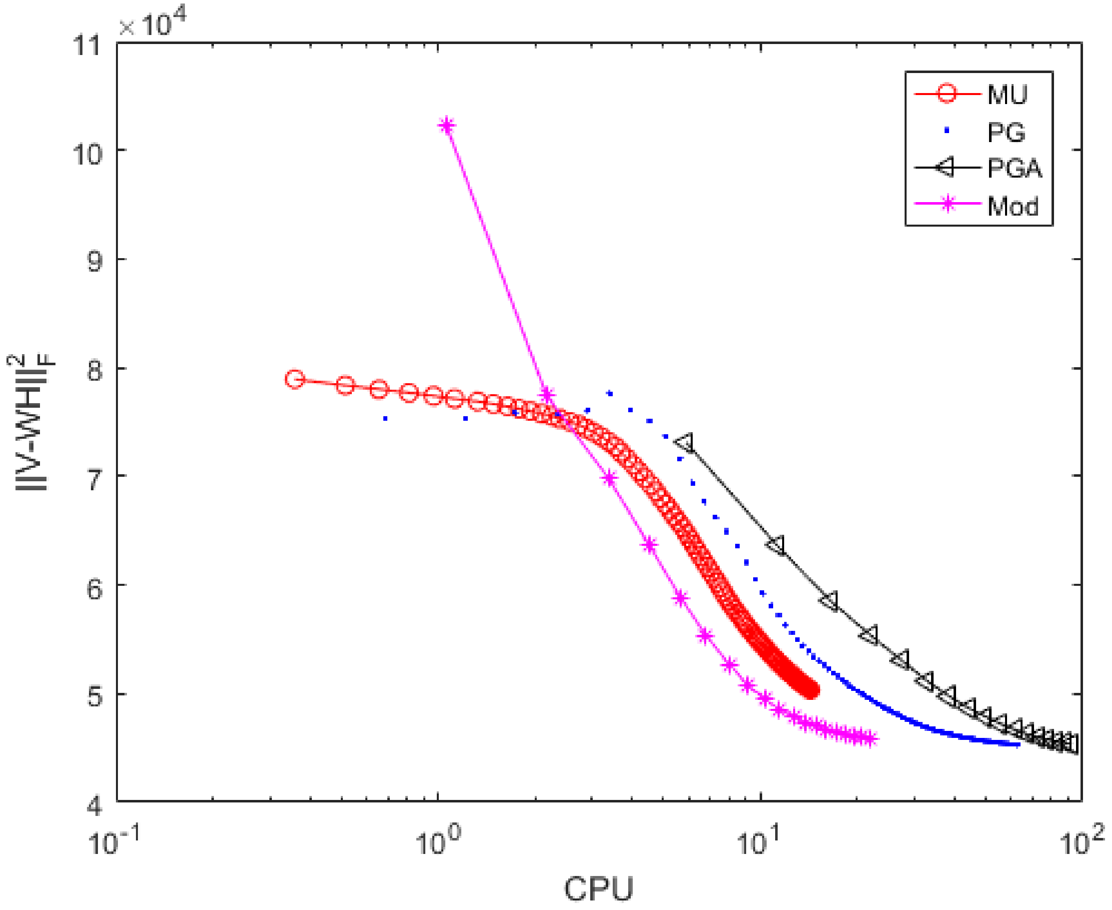
Comparison of the iterative methods for ORL facedata problem.

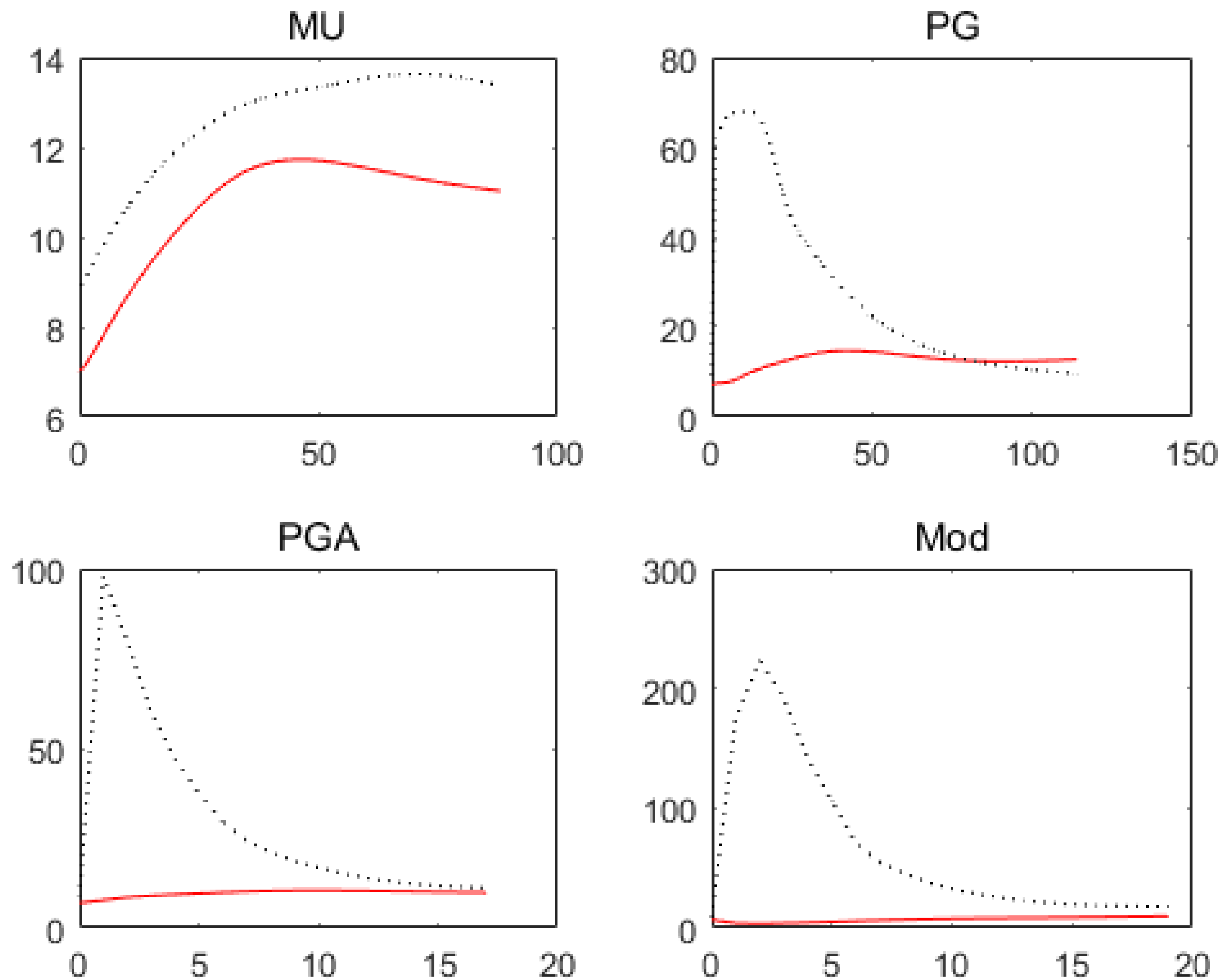
Methods	Iterations	$f(W, H)$	CPU
MU	88	50346.85	14.28
PG	114	45296.18	63.02
PGA	17	45372.30	92.58
Mod	19	45900.87	21.83

ORL facedata problem: iterations

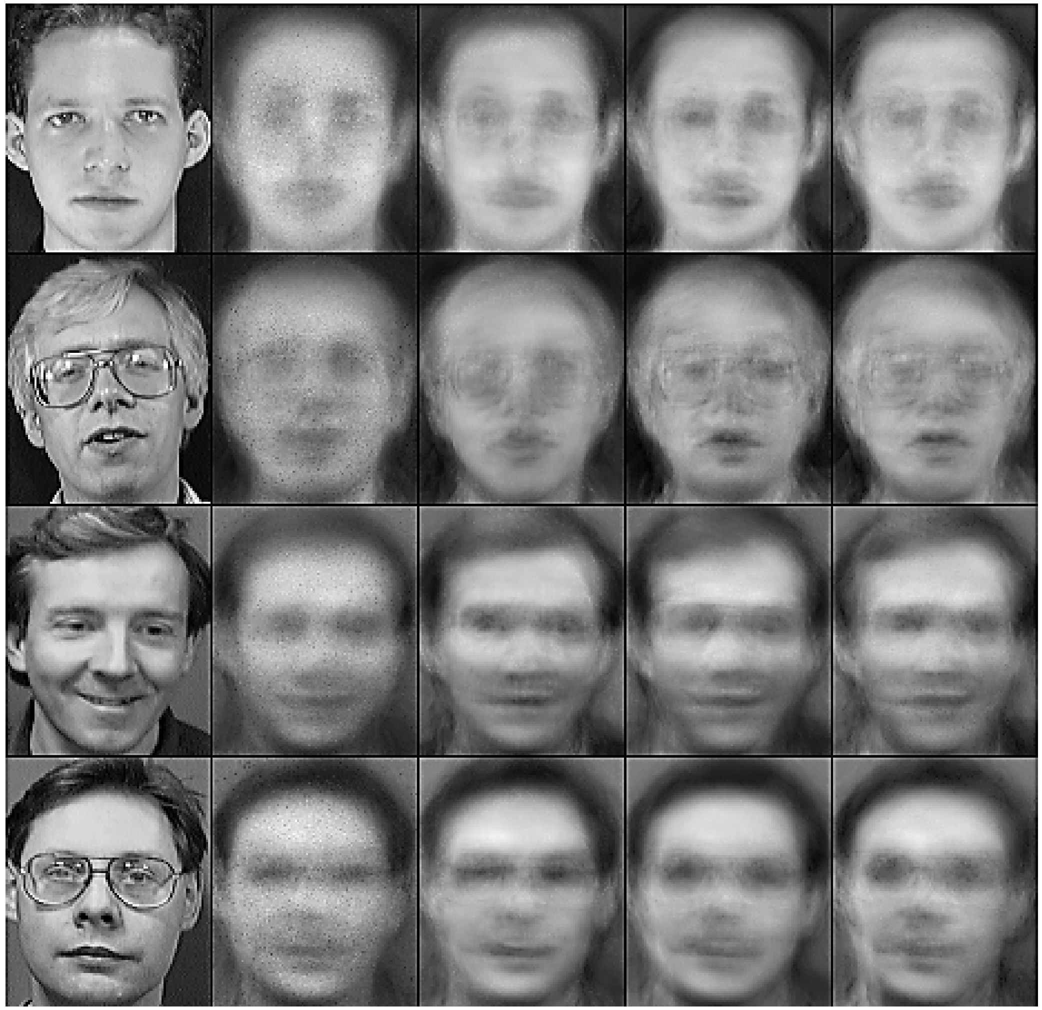


ORL facedata problem: computational time





Iterations vs. condition numbers of W^k (red line) and H^k (black dot).



From left column to right column: original images, MU, PG, PGA, Mod

Concluding Remarks

Alternating nonnegative least squares method

$$\downarrow$$

Modulus method for the subproblems

$$\downarrow$$

- Competitive among the previous methods
- The optimal modulus-type inner outer iteration method can be further exploited

Future Work: Sparse NMF

• Add penalty terms to the NMF objective function (Hoyer, 02')

$$\min \frac{1}{2}\|V - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2,$$

where α and β are positive parameters.

• Minimize the (generalized) Kullback-Leibler divergence between V and WH (Lee and Seung, 99')

$$\min \sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right)$$

Future Work: Sparse NMF

Sparsity constraint with Frobenius norm:

$$\min \|V - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \|H\|_F^2,$$

Alternating nonnegative least squares method

$$\begin{aligned}
\min \|V - W^k H\|_F^2 + \beta \|H\|_F^2 &= \min \left\| \begin{bmatrix} V \\ 0 \end{bmatrix} - \begin{bmatrix} W^k \\ \sqrt{\beta} I \end{bmatrix} H \right\|_F^2 \\
&:= \min \|\bar{V} - \bar{W}^k H\|_F^2 \\
\min \|V - WH^{k+1}\|_F^2 + \alpha \|W\|_F^2 &= \min \left\| \begin{bmatrix} V & 0 \end{bmatrix} - W \begin{bmatrix} H^{k+1} & \sqrt{\alpha} I \end{bmatrix} \right\|_F^2 \\
&:= \min \|\tilde{V} - W\tilde{H}^{k+1}\|_F^2,
\end{aligned}$$

where I is an identity matrix.

Future Work: Sparse NMF

Sparsity constraint with L1-norm:

$$\min \|V - WH\|_F^2 + \alpha \sum_{i=1}^{m} \|w_i\|_1^2 + \beta \sum_{j=1}^{n} \|h_j\|_1^2,$$

where w_i^T and h_j are ith row vector of W and jth column vector of W, respectively. Alternating nonnegative least squares method

$$\begin{aligned}
\min \|V - W^k H\|_F^2 + \beta \sum_{j=1}^{n} \|h_j\|_1^2 &= \min \left\| \begin{bmatrix} V \\ 0 \end{bmatrix} - \begin{bmatrix} W^k \\ \sqrt{\beta}\mathbf{e}^\mathsf{T} \end{bmatrix} H \right\|_F^2 \\
&:= \min \|\bar{V} - \bar{W}^k H\|_F^2 \\
\min \|V - WH^{k+1}\|_F^2 + \alpha \sum_{i=1}^{m} \|w_i\|_1^2 &= \min \left\| \begin{bmatrix} V & 0 \end{bmatrix} - W \begin{bmatrix} H^{k+1} & \sqrt{\alpha}\mathbf{e} \end{bmatrix} \right\|_F^2 \\
&:= \min \|\tilde{V} - W\tilde{H}^{k+1}\|_F^2,
\end{aligned}$$

where $\mathbf{e}$ is a column vector with all components equal to one.

Thank You!