Supervised Matrix Factorization: Local Landscape Analysis and Applications

Joowon Lee[1], **Hanbaek Lyu**[2], and **Weixin Yao**[3]

1: Department of Statistics, University of Wisconsin – Madison, WI, USA
2: Department of Mathematics, University of Wisconsin – Madison, WI, USA
3: Department of Statistics , University of California, Riverside, CA USA

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Abstract

Supervised matrix factorization (SMF) is a classical machine learning method that seeks low-dimensional feature extraction and classification tasks at the same time. Training an SMF model involves solving a non-convex and factor-wise constrained optimization problem with at least three blocks of parameters. Due to the high non-convexity and constraints, theoretical understanding of the optimization landscape of SMF has been limited. In this paper, we provide an extensive local landscape analysis for SMF and derive several theoretical and practical applications. Analyzing diagonal blocks of the Hessian naturally leads to a block coordinate descent (BCD) algorithm with adaptive step sizes. We provide global convergence and iteration complexity guarantees for this algorithm. Full Hessian analysis gives minimum L^2-regularization to guarantee local strong convexity and robustness of parameters. We establish a local estimation guarantee under a statistical SMF model. We also propose a novel GPU-friendly neural implementation of the BCD algorithm and validate our theoretical findings through numerical experiments. Our work contributes to a deeper understanding of SMF optimization, offering insights into the optimization landscape and providing practical solutions to enhance its performance.

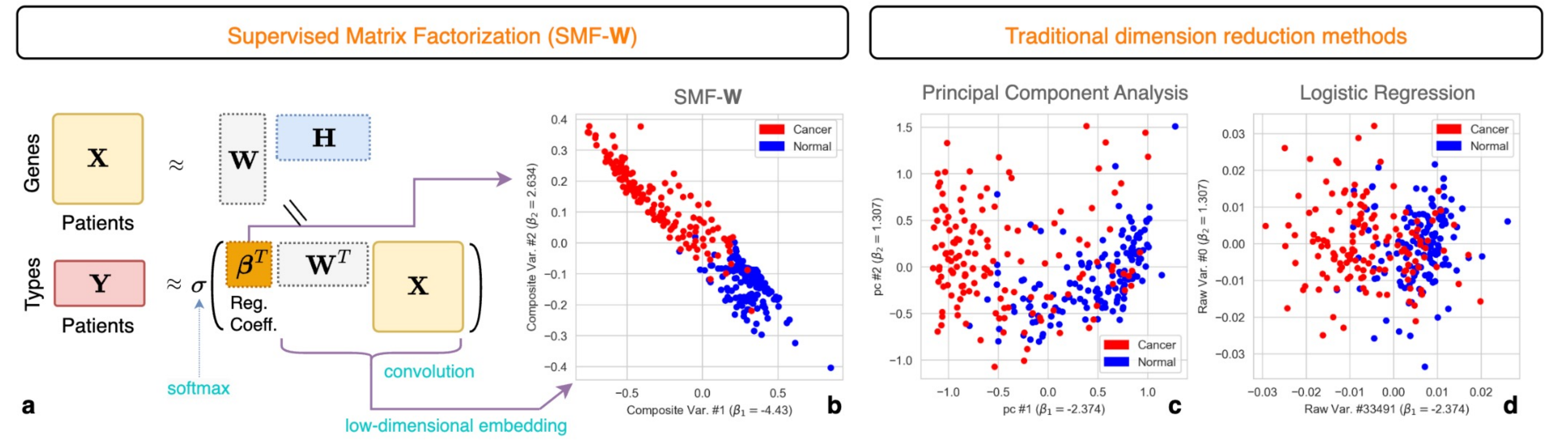


Figure 1. (**a**) Overall scheme of Supervised Matrix Factorization (specifically, SMF-$\mathbf{W}$ with rank $r = 2$). The columns of $\mathbf{W}$ serve as 'composite variables' or 'filters', whose association with the labels is given by the regression coefficients in $\boldsymbol{\beta}$. Taking convolution of the raw data matrix $\mathbf{W}$ with $\mathbf{W}$ gives a supervised dimension reduction, as illustrated in **b** for a 35,982-dimensional gene microarray data for breast cancer patients. Similar dimension reduction results obtained by (**c**) principal component analysis along with logistic regression and (**d**) logistic regression to select the two most highly associated raw variables show less clear separation.

1. Objective and Model Formulation

- Labeled data $(y_i, \mathbf{x}_i)$, $i = 1, \dots, n$:
  - y_i = binary label (e.g., 1="Cancer", 0="Normal")
  - $\mathbf{x}_i$ = high-dimensional feature vector (e.g., gene expression data)
- **Dimension reduction + Classification** at the same time?
- Key difficulty:

Best direction for dimension reduction ≠ Good label separation

- **Supervised Matrix Factorization** – Joint optimization formulation

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta}} f(\mathbf{W},\mathbf{H},\boldsymbol{\beta}) := \underbrace{\sum_{i=1}^{n} \ell(y_i, \boldsymbol{\beta}^T\mathbf{W}^T\mathbf{x}_i)}_{\text{Classification loss}} + \xi \underbrace{\|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2}_{\text{Dimension reduction loss}}$$

Negative log-likelihood under logistic model

$$\ell(y,a) = \log(1+\exp(a)) - \mathbf{1}_{\{y=1\}} a$$

- ○ $\mathbf{W} \in \mathbb{R}_{\geq 0}^{p\times r}$: high-dim to low-dim filter (supervised PCs)
- ○ $\boldsymbol{\beta} \in \mathbb{R}^{1\times r}$: regression coefficients
- ○ $\xi \geq 0$: trade-off parameter ($\xi = 0 \rightarrow$ 2-layer NN)

2. How the model works

- The filter matrix $\mathbf{W}$ is learned to serve double purposes:
  - ○ Input for Classification : $\mathbf{W}^T\mathbf{x}_i = r$ – dim compressed feature
  - ○ Feature reconstruction : $\mathbf{X} \approx \mathbf{W}\mathbf{H}$ for some $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r\times n}$
- Once the filter $\mathbf{W}$ and reg. coefficients $\boldsymbol{\beta}$ are learned:

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\boldsymbol{a}_i)}{1+\exp(\boldsymbol{a}_i)}, \quad \text{where } \boldsymbol{a}_i = \boldsymbol{\beta}^T\mathbf{W}^T\mathbf{x}_i$$

3. Local Optimization Landscape: Diagonal

- To sketch the idea, consider the matrix factorization loss only:

$$f(\mathbf{W},\mathbf{H}) = \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2$$

$$\nabla^2 f = \begin{array}{c c} & \begin{array}{cc} \mathrm{vec}(\mathbf{W})^T & \mathrm{vec}(\mathbf{H})^T \end{array} \\ \begin{array}{c} \mathrm{vec}(\mathbf{W}) \\ \mathrm{vec}(\mathbf{H}) \end{array} & \begin{bmatrix} \mathbf{H}\mathbf{H}^T \otimes \mathbf{I}_p & A_{12} \\ A_{12}^T & \mathbf{I}_n \otimes \mathbf{W}^T\mathbf{W} \end{bmatrix} \end{array}$$

$\lambda_{max}(\nabla^2 f)$ Unbounded ☹ → PGD very sensitive on step size

$\lambda_{max}(\nabla^2_{\mathbf{W}} f(\cdot,\mathbf{H})) = \lambda_{\max}(\mathbf{H}\mathbf{H}^T)$ Bounded for fixed **H !!** ☺

$\lambda_{max}(\nabla^2_{H} f(\mathbf{W},\cdot)) = \lambda_{\max}(\mathbf{W}^T\mathbf{W})$ Bounded for fixed **W !!** ☺

Adaptive BCD:

$$\mathbf{W} \leftarrow \Pi\left(\mathbf{W} - \frac{1}{\lambda_{\max}(\mathbf{H}\mathbf{H}^T)+\varepsilon}(\mathbf{W}\mathbf{H}-\mathbf{X})\mathbf{H}^T\right),$$

$$\mathbf{H} \leftarrow \Pi'\left(\mathbf{H} - \frac{1}{\lambda_{\max}(\mathbf{W}\mathbf{W}^T)+\varepsilon}\mathbf{W}^T(\mathbf{W}\mathbf{H}-\mathbf{X})\right)$$

Largest Eval of *diagonal blocks* of the Hessian

≪ Largest Eval of the entire Hessian

- We derive Adaptive BCD for SMF similarly

$$\begin{array}{c c} & \begin{array}{cccc} \mathrm{vec}(\mathbf{W})^T & \mathrm{vec}(\mathbf{H})^T & \mathrm{vec}(\boldsymbol{\beta})^T & \mathrm{vec}(\boldsymbol{\Gamma})^T \end{array} \\ \begin{array}{c} \mathrm{vec}(\mathbf{W}) \\ \mathrm{vec}(\mathbf{H}) \\ \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vec}(\boldsymbol{\Gamma}) \end{array} & \begin{bmatrix} A_{11} & A_{12} & A_{13} & O \\ A_{21} & A_{22} & O & O \\ A_{31} & O & A_{33} & A_{34} \\ O & O & A_{43} & A_{44} \end{bmatrix} \end{array}$$

4. Local Optimization Landscape: Off-diagonal

$$\|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 + \lambda_1\|\mathbf{W}\|_F^2 + \lambda_2\|\mathbf{H}\|_F^2$$

- Smallest L_2-reg. to ensure *local strong convexity*?

(→ Robust parameter estimation)

- ○ Smallest L_2-reg. for block diagonal dominance in $\nabla^2 f$

$$\lambda_{min}(\mathbf{H}\mathbf{H}^T) + \lambda_1 - \|A_{12}\|_2 > 0$$

$$\lambda_{min}(\mathbf{W}^T\mathbf{W}) + \lambda_2 - \|A_{12}\|_2 > 0$$

- ○ High-dim regime: $p \gg n$ → No need to reg. **H**!!

$$\lambda_{min}(\mathbf{W}^T\mathbf{W}) = \Theta(rp) \gg \Theta(r\sqrt{pn}) = \|A_{12}\|_2$$

→ $\mathbf{H}^{opt}$ can be recovered robustly despite non-convexity

- ○ Large-sample regime: $p \ll n$ → No need to reg. **W**!!

$$\lambda_{min}(\mathbf{H}\mathbf{H}^T) = \Theta(rn) \gg \Theta(r\sqrt{pn}) = \|A_{12}\|_2$$

Theorem 4.5. *(Regularized local consistency) Consider the generative SMF-$\mathbf{W}$ model (16). Assume that Assumptions 4.1 and 4.2 hold. Suppose $\rho := \min_{1\leq i\leq 4}(\lambda_i - \lambda_{i\star}) > 0$.*

Suppose $\Lambda_1 > 0$, $\lambda_1 = 0$, and $\sigma \ll 1$ (resp., $\Lambda_2 > 0$ and $\lambda_2 = 0$). Fix $\varepsilon > 0$. Then there exists a constant $C > 0$ such that with probability at least $1-\varepsilon$, $\mathcal{L}$ in (17) is minimized locally at some $(\hat{\mathbf{W}}, \hat{\theta}, \hat{\boldsymbol{\lambda}})$ (resp., $(\hat{\mathbf{H}}, \hat{\theta}, \hat{\boldsymbol{\lambda}})$) with

$$\|\hat{\mathbf{W}} - \mathbf{W}_\star\| \leq C/\sqrt{n} \text{ (resp., } \|\hat{\mathbf{H}} - \mathbf{H}_\star\| \leq C/\sqrt{n}) \quad (18)$$

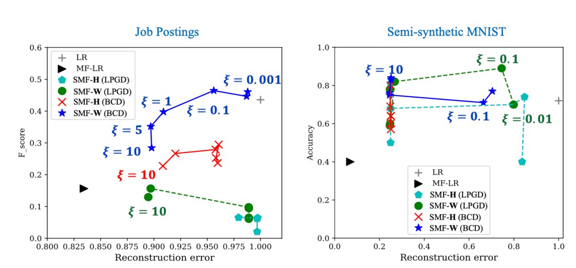
$$\|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_\star\| \leq C/\sqrt{n}$$

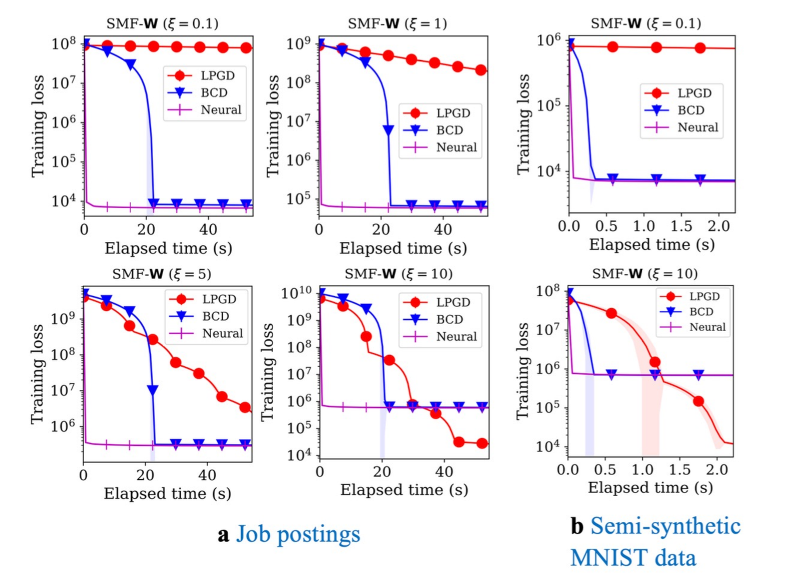
$$\|\hat{\theta} - \theta_\star\|_F \leq Cn^{-1/2}\left(1 + \frac{3\max\{\lambda_2,\lambda_3,\lambda_4\}}{\rho}\|\theta_\star\|_F\right),$$

where $\theta' := (\mathbf{H}', \boldsymbol{\beta}', \boldsymbol{\Gamma}')$, $\theta_\star := (\mathbf{H}_\star, \boldsymbol{\beta}_\star, \boldsymbol{\Gamma}_\star)$ (resp., $\theta' := (\mathbf{W}', \boldsymbol{\beta}', \boldsymbol{\Gamma}')$, $\theta_\star := (\mathbf{W}_\star, \boldsymbol{\beta}_\star, \boldsymbol{\Gamma}_\star)$) and $\|\theta_\star\|_F$ is assumed to be sufficiently small.

5. Experiments

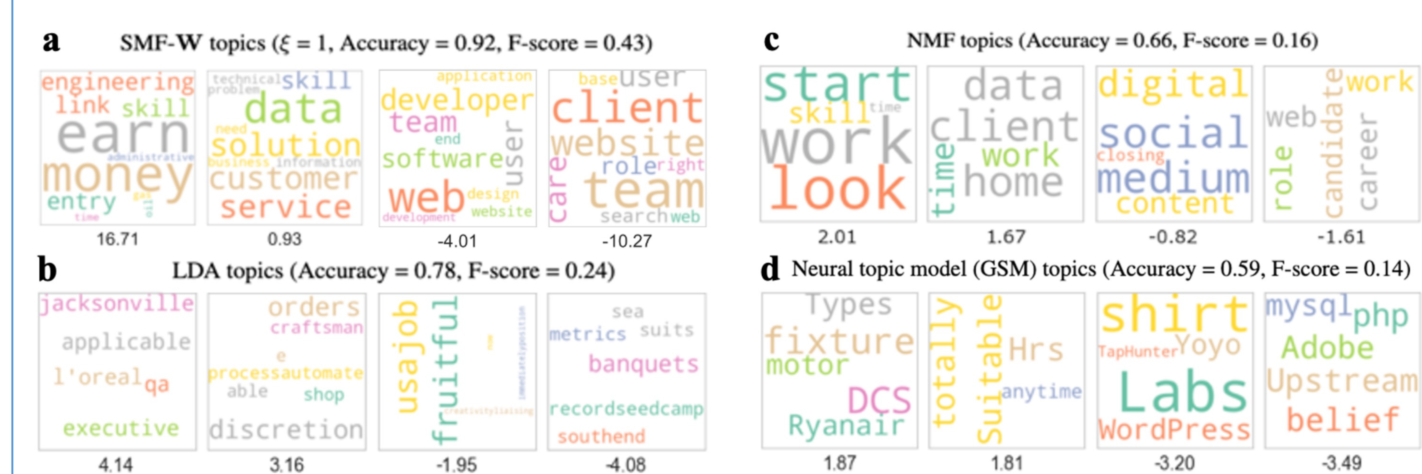
Methods	Pancreatic	Breast
SMF-**W** (BCD)	0.869 (0.02)	**0.924 (0.01)**
SMF-**H** (BCD)	0.823 (0.06)	0.880 (0.02)
SMF-**W** (Neural)	0.854 (0.04)	0.881 (0.02)
SMF-**W** (LPGD)	0.869 (0.02)	0.894 (0.02)
SMF-**H** (LPGD)	**0.885 (0.07)**	0.875 (0.01)
PCA-LR	0.747 (0.13)	0.454 (0.27)
CNN	0.769 (0.07)	0.854 (0.06)
FFNN	0.816 (0.04)	0.890 (0.02)
Naive Bayes	0.815 (0.07)	0.810 (0.02)
SVM	0.746 (0.09)	0.866 (0.02)
Random Forest	0.815 (0.06)	0.844 (0.02)





a Job postings **b** Semi-synthetic MNIST data

Figure 3. Plots of training loss vs. elapsed time at different ξ values for fitting SMF-$\mathbf{W}$ using Algorithm 1 (BCD), the neural implementation in Figure 2 (Neural), and low-rank projected gradient descent (LPGD) in (Lee et al., 2023). Shaded regions indicate one standard deviation across 10 runs.



6. References

Lee, J., Lyu, H., and Yao, W. Exponentially convergent algorithms for supervised matarix factorization. *NeurIPS 2023*

7. Acknowledgements

NSF DMS-2206296, DMS—010035, DMS-2210272