SWIFT: Scalable Wasserstein Factorization for Sparse Nonnegative Tensors

Ardavan Afshar[1] Kejing Yin[2] Sherry Yan[3] Cheng Qian[4] Joyce C. Ho[5] Haesun Park[1] Jimeng Sun[6]

[1] Georgia Institute of Technology [2] Hong Kong Baptist University [3] Sutter Health [4] IQVIA [5] Emory University [6] University of Illinois at Urbana-Champaign

Contributions

1. **Defining Wasserstein Tensor Distance.**
   The first work that defines Wasserstein distance for tensors.

2. **Formulating Wasserstein Tensor Factorization.**
   SWIFT model minimizes the Wasserstein distance between the input and its CP reconstructions.

3. **Efficiently Solving Wasserstein TF.**
   It achieves 921x speed up over a naive implementation.

Motivations & Preliminaries

Existing tensor factorization models assume certain distributions of input, e.g.,

Gaussian distribution: $\min_{\hat{\mathcal{X}}} \ \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2$ ← *MSE loss*

Poisson distribution: $\min_{\hat{\mathcal{X}}} \ \hat{\mathcal{X}} - \mathcal{X} * \log(\hat{\mathcal{X}})$ ← *KL divergence*

Bernoulli distribution: $\min_{\hat{\mathcal{X}}} \ \log(1 + e^{\hat{\mathcal{X}}}) - \mathcal{X} * \hat{\mathcal{X}}$ ← *logit loss*

However, the distribution of input tensor is often **complex and unknown**.

Wasserstein distance is a potentially better metric:

Definition (Entropy regularized OT problem)

The entropy regularized OT problem is defined as:

$$W_V(\mathbf{a}, \mathbf{b}) = \underset{\mathbf{T} \in U(\mathbf{a},\mathbf{b})}{\text{minimize}} \quad \langle \mathbf{C}, \mathbf{T} \rangle - \frac{1}{\rho} E(\mathbf{T}),$$

where $E(\mathbf{T}) = -\sum_{i,j=1}^{M,N} t_{ij} \log(t_{ij})$ is the entropy of $\mathbf{T}$.

☹ **Not directly applicable to tensor factorization:**

1) It is not defined for tensor input;
2) It requires solving expensive OT problems for many times.

Wasserstein Matrix and Tensor Distances

Wasserstein Matrix Distance: sum W_V over their vectors:

Definition (Wasserstein Matrix Distance)

Given a cost matrix $\mathbf{C} \in \mathbb{R}_+^{M \times M}$, the Wasserstein distance between two matrices $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_P] \in \mathbb{R}_+^{M \times P}$ and $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_P] \in \mathbb{R}_+^{M \times P}$ is denoted by $W_M(\mathbf{A}, \mathbf{B})$, and given by:

$$\boxed{W_M(\mathbf{A}, \mathbf{B}) = \sum_{p=1}^{P} W_V(\mathbf{a}_p, \mathbf{b}_p)} = \underset{\overline{\mathbf{T}} \in U(\mathbf{A},\mathbf{B})}{\text{minimize}} \langle \overline{\mathbf{C}}, \overline{\mathbf{T}} \rangle - \frac{1}{\rho} E(\overline{\mathbf{T}}), \tag{3}$$

where $\overline{\mathbf{C}} = [\mathbf{C}, ..., \mathbf{C}]$ and $\overline{\mathbf{T}} = [\mathbf{T}_1, ...\mathbf{T}_p, ..., \mathbf{T}_P]$. $U(\mathbf{A}, \mathbf{B}) = \left\{ \overline{\mathbf{T}} \in \mathbb{R}_+^{M \times MP} \mid \Delta(\overline{\mathbf{T}}) = \mathbf{A}, \Psi(\overline{\mathbf{T}}) = \mathbf{B} \right\}$, $\Delta(\overline{\mathbf{T}}) = [\mathbf{T}_1 \mathbf{1}_M, ..., \mathbf{T}_P \mathbf{1}_M] = \overline{\mathbf{T}}(\mathbf{I}_P \otimes \mathbf{1}_M)$, and $\Psi(\overline{\mathbf{T}}) = [\mathbf{T}_1^T \mathbf{1}_M, ..., \mathbf{T}_P^T \mathbf{1}_M]$.

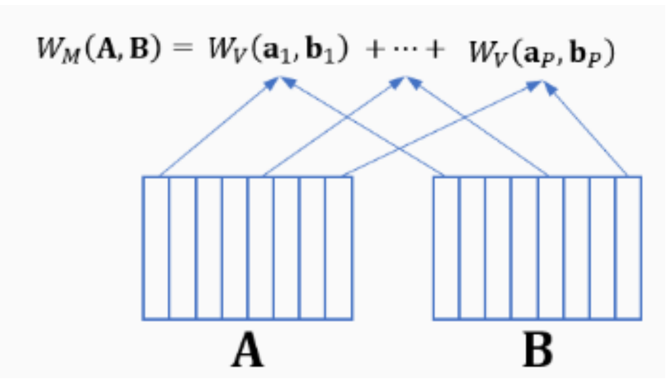
Wasserstein Tensor Distance: sum W_M over their matricizations:

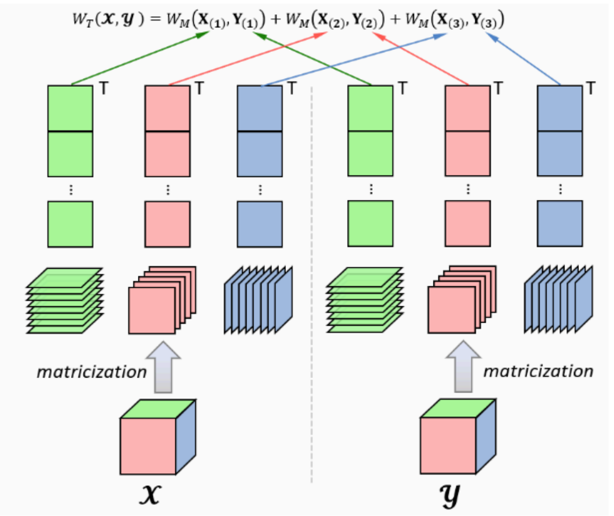
Definition (Wasserstein Tensor Distance)

The Wasserstein distance between N-th order tensor $\mathcal{X} \in \mathbb{R}_+^{I_1 \times \ldots \times I_N}$ and its reconstruction $\hat{\mathcal{X}} \in \mathbb{R}_+^{I_1 \times \ldots \times I_N}$ is denoted by $W_T(\hat{\mathcal{X}}, \mathcal{X})$:

$$\boxed{W_T(\hat{\mathcal{X}}, \mathcal{X}) = \sum_{n=1}^{N} W_M\left(\widehat{\mathbf{X}}_{(n)}, \mathbf{X}_{(n)}\right)} \equiv \sum_{n=1}^{N} \left\{ \underset{\overline{\mathbf{T}}_n \in U\left(\widehat{\mathbf{X}}_{(n)}, \mathbf{X}_{(n)}\right)}{\text{minimize}} \quad \langle \overline{\mathbf{C}}_n, \overline{\mathbf{T}}_n \rangle - \frac{1}{\rho} E(\overline{\mathbf{T}}_n) \right\}, \tag{4}$$

where $\mathbf{X}_{(n)} \in \mathbb{R}_+^{I_n \times I_{(-n)}}$ is the n-th mode matricization of $\mathcal{X}$, $\overline{\mathbf{C}}_n = [\mathbf{C}_n, \mathbf{C}_n, ..., \mathbf{C}_n] \in \mathbb{R}_+^{I_n \times I_n I_{(-n)}}$, and $\overline{\mathbf{T}}_n = [\mathbf{T}_{n1}, ..., \mathbf{T}_{nj}, ..., \mathbf{T}_{nI_{(-n)}}] \in \mathbb{R}_+^{I_n \times I_n I_{(-n)}}$. $\mathbf{T}_{nj} \in \mathbb{R}_+^{I_n \times I_n}$ is the transport matrix between the columns $\widehat{\mathbf{X}}_{(n)}(:, j) \in \mathbb{R}_+^{I_n}$ and $\mathbf{X}_{(n)}(:, j) \in \mathbb{R}_+^{I_n}$.





Wasserstein Tensor Factorization

Optimization problem:

Constraint Relaxation using the generalized KL-divergence (5)

$$\underset{\{\mathbf{A}_n \geq 0, \overline{\mathbf{T}}_n\}_{n=1}^N}{\text{minimize}} \quad \sum_{n=1}^{N} \left(\underbrace{\langle \overline{\mathbf{C}}_n, \overline{\mathbf{T}}_n \rangle - \frac{1}{\rho} E(\overline{\mathbf{T}}_n)}_{\text{Part } P_1} + \lambda \left(\underbrace{KL(\Delta(\overline{\mathbf{T}}_n) || \mathbf{A}_n (\mathbf{A}_{\odot}^{(-n)})^T)}_{\text{Part } P_2} + \underbrace{KL(\Psi(\overline{\mathbf{T}}_n) || \mathbf{X}_{(n)})}_{\text{Part } P_3} \right) \right)$$

SWIFT Learning Algorithm

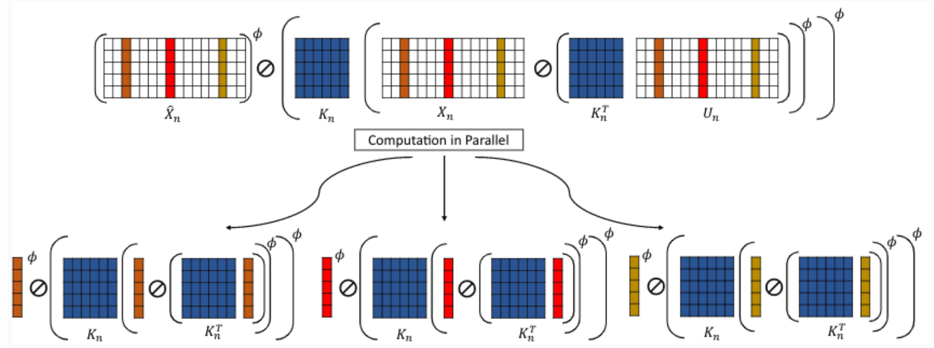
1. We avoid computing OT by $\mathbf{T}_{nj}^* \mathbf{1} = \text{diag}(\mathbf{u}_j) \mathbf{K}_n \mathbf{v}_j = \mathbf{u}_j * (\mathbf{K}_n \mathbf{v}_j)$

Proposition 2

$$\Delta(\overline{\mathbf{T}}_n) = [\mathbf{T}_{n1}\mathbf{1}, ..., \mathbf{T}_{nj}\mathbf{1}, ..., \mathbf{T}_{nI_{(-n)}}\mathbf{1}] = \mathbf{U}_n * (\mathbf{K}_n \mathbf{V}_n) \tag{6}$$

minimizes (5), where $\mathbf{K}_n = e^{(-\rho \mathbf{C}_n - 1)} \in \mathbb{R}_+^{I_n \times I_n}$, $\mathbf{U}_n = (\widehat{\mathbf{X}}_{(n)})^{\Phi} \oslash \left(\mathbf{K}_n (\mathbf{X}_{(n)} \oslash (\mathbf{K}_n^T \mathbf{U}_n))^{\Phi}\right)^{\Phi}$, $\mathbf{V}_n = (\mathbf{X}_{(n)} \oslash (\mathbf{K}_n^T \mathbf{U}_n))^{\Phi}$, $\Phi = \frac{\lambda\rho}{\lambda\rho+1}$, and $\oslash$ indicates element-wise division.

2. We explore sparsity structure of input: All-zero columns in $\mathbf{X}_{(n)}$ are ignored.



SWIFT explores sparsity structure in input data $\mathbf{X}_{(n)}$ and drops zero values columns.

3. We introduce efficient rearrangement for updating factor matrices to decouple $\mathbf{A}_{(n)}$ and Khatri-Rao product.

Define: $\Pi(\mathbf{A}_i (\mathbf{A}_{\odot}^{(-i)})^T, n) = \mathbf{A}_n (\mathbf{A}_{\odot}^{(-n)})^T \in \mathbb{R}_+^{I_n \times I_{(-n)}} \quad \forall \, i \neq n.$

Rearranged subproblem for $\mathbf{A}_{(n)}$:

$$\underset{\mathbf{A}_n \geq 0}{\text{minimize}} \quad KL\left(\begin{bmatrix} \Pi(\Delta(\overline{\mathbf{T}}_1), n) \\ \cdots \\ \Pi(\Delta(\overline{\mathbf{T}}_i), n) \\ \cdots \\ \Pi(\Delta(\overline{\mathbf{T}}_N), n) \end{bmatrix} \middle\| \begin{bmatrix} \mathbf{A}_n (\mathbf{A}_{\odot}^{(-n)})^T \\ \cdots \\ \mathbf{A}_n (\mathbf{A}_{\odot}^{(-n)})^T \\ \cdots \\ \mathbf{A}_n (\mathbf{A}_{\odot}^{(-n)})^T \end{bmatrix} \right)$$

Datasets and Baselines

1. **BBC New**: 400 article x 100 words x 100 words
   task: article category classification, evaluate by accuracy.

2. **Sutter**: 1000 patients x 100 diagnoses x 100 medications
   task: heart failure onset, evaluate by PR-AUC.

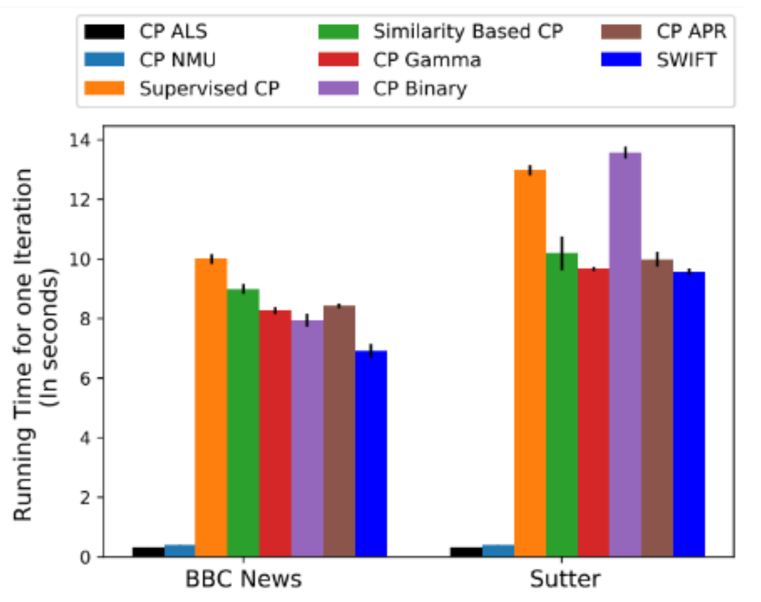
Baselines:

1. MSE Loss (*Gaussian*): CP-ALS; CP-NMU; Supervised CP; Similarity based CP;
2. Gamma Loss (*Gamma distribution*): CP-Continuous;
3. Log Loss (*Bernoulli distribution*): CP-Binary;
4. KL Loss (*Poisson distribution*): CP-APR.

Results

		R=5	R=10	R=20	R=30	R=40
BBC News Dataset	CP-ALS	.521 ± .033	.571 ± .072	.675 ± .063	.671 ± .028	.671 ± .040
	CP-NMU	.484 ± .039	.493 ± .048	.581 ± .064	.600 ± .050	.650 ± .031
	Supervised CP	.506 ± .051	.625 ± .073	.631 ± .050	.665 ± .024	.662± .012
	Similarity Based CP	.518 ± .032	.648 ± .043	.638 ± .021	.662 ± .034	.673 ± .043
	CP-Continuous	.403 ± .051	.481 ± .056	.528 ± .022	.559 ± .024	.543 ± .043
	CP-Binary	.746 ± .058	.743 ± .027	.737 ± .008	.756 ± .062	.743 ± .044
	CP-APR	.675 ± .059	.768 ± .033	.753 ± .035	.743 ± .033	.746 ± .043
	SWIFT	**.759 ± .013**	**.781 ± .013**	**.803 ± .010**	**.815 ± .005**	**.818 ± .022**
Sutter Data	CP-ALS	.327 ± .072	.333 ± .064	.311 ± .068	.306 ± .065	.332 ± .098
	CP-NMU	.300 ± .054	.294 ± .064	.325 ± .085	.344 ± .068	.302 ± .071
	Supervised CP	.301 ± .044	.305 ± .036	.309 ± .054	.291 ± .037	.293 ± .051
	Similarity Based CP	.304 ± .042	.315 ± .041	.319 ± .063	.296 ± .041	.303 ± .032
	CP-Continuous	.252 ± .059	.237 ± .043	.263 ± .065	.244 ± .053	.256 ± .077
	CP-Binary	.301 ± .061	.325 ± .079	.328 ± .080	.267 ± .074	.296 ± .063
	CP-APR	.305 ± .075	.301 ± .068	.290 ± .052	.313 ± .082	.304 ± .086
	SWIFT	**.364 ± .063**	**.350 ± .031**	**.350 ± .040**	**.369 ± .066**	**.374 ± .044**

Classification Performance



Atrial Fibrillation (Weight= 21.93)
Dx-Essential hypertension [98.]
Dx-Disorders of lipid metabolism [53.]
Dx-Cardiac dysrhythmias [106.]
Rx-Calcium Channel Blockers
Rx-Alpha-Beta Blockers
Rx-Angiotensin II Receptor Antagonists
Cardiometablic Disease (Weight= 19.58)
Dx-Diabetes mellitus without complication [49.]
Dx-Essential hypertension [98.]
Dx-Disorders of lipid metabolism [53.]
Rx-Diagnostic Tests
Rx-Biguanides
Rx-Diabetic Supplies
Mental Disorder (Weight= -16.22)
Dx-Anxiety disorders [651]
Dx-Menopausal disorders [173.]
Dx-Depressive disorders [6572]
Rx-Benzodiazepines
Rx-Selective Serotonin Reuptake Inhibitors (SSRIs)
Rx-Serotonin Modulators

Scalable and Interpretable

References

1. Tamara G Kolda and Brett W Bader. "Tensor decompositions and applications". In: SIAM Review (2009).
2. J Carroll and J Chang. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition". In:Psychometrika (1970).
3. E Chi and T Kolda. "On tensors, sparsity, and nonnegative factorizations". In: SIAM Journal on Matrix Analysis and Applications (2012).
4. D Hong, T Kolda, and J Duersch. "Generalized canonical polyadic tensor decomposition". In: SIAM Review (2020).
5. Gabriel Peyŕ e, Marco Cuturi, et al. "Computational optimal transport". In: Foundations and Trends® in Machine Learning (2019).
6. Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: Advances in Neural Information Processing Systems. 2013.
7. Richard Sinkhorn and Paul Knopp. "Concerning nonnegative matrices and doubly stochastic matrices". In: Pacific Journal of Mathematics (1967).
8. Daniel D Lee and H Sebastian Seung. "Algorithms for non-negative matrix factorization". In: Advances in Neural Information Processing Systems. 2001.
9. erek Greene and Ṕ adraig Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering". In: InternationalConference on Machine learning. 2006.

Acknowledgments: This work is in part supported by National Science Foundation award SCH-2014438, IIS-1418511, CCF-1533768, IIS-1838042, the National Institute of Health award NIH R01 1R01NS107291-01 and R56HL138415.

Please check out our paper for more details: https://arxiv.org/pdf/2010.04081.pdf
Thank you for your interest in our work.