A Contrastive Estimation with the Picard iteration

Letting $\beta = |\mathcal{A}^+| - |\mathcal{A}^-| \geq 0$ and writing $\boldsymbol{U}_A$ as the $M \times |A|$ indicator matrix such that $\boldsymbol{L}_A = \boldsymbol{U}_A^\top \boldsymbol{L} \boldsymbol{U}_A$, we have

$$\phi(\boldsymbol{L}) \propto \underbrace{-\beta \log\det(\boldsymbol{I}+\boldsymbol{X}) + \sum_{A\in\mathcal{A}^+} \log\det(\boldsymbol{U}_A^\top \boldsymbol{X}^{-1} \boldsymbol{U}_A)}_{f \text{ convex}}$$

$$+ \underbrace{\beta \log\det(\boldsymbol{X}) - \sum_{A\in\mathcal{A}^-} \log\det(\boldsymbol{U}_A^\top \boldsymbol{X}^{-1} \boldsymbol{U}_A)}_{g \text{ concave}}$$

where the convexity/concavity results follow immediately from [Mariet and Sra, 2015, Lemma 2.3]. Then, the update rule $\nabla f(\boldsymbol{L}_{k+1}) = -\nabla g(\boldsymbol{L}_k)$ requires

$$\beta \boldsymbol{L}_{k+1} + \sum_{A\in\mathcal{A}^-} \boldsymbol{L}_{k+1}\boldsymbol{U}_A(\boldsymbol{U}_A^\top \boldsymbol{L}_{k+1}\boldsymbol{U}_A)^{-1}\boldsymbol{U}_A^\top \boldsymbol{L}_{k+1}$$

$$\leftarrow \beta(\boldsymbol{I}+\boldsymbol{L}_k^{-1})^{-1} + \sum_{A\in\mathcal{A}^+} \boldsymbol{L}_k\boldsymbol{U}_A(\boldsymbol{U}_A^\top \boldsymbol{L}_k\boldsymbol{U}_A)^{-1}\boldsymbol{U}_A^\top \boldsymbol{L}_k$$

which cannot be evaluated due to the $\sum_{A\in\mathcal{A}^-}$ term.

B Amazon Baby registries experiments

B.1 Amazon Baby Registries description

Table 3: Description of the Amazon Baby registries dataset.

REGISTRY	M	TRAIN SIZE	TEST SIZE
HEALTH	62	5278	1320
BATH	100	5510	1377
APPAREL	100	6482	1620
BEDDING	100	7119	1780
DIAPER	100	8403	2101
GEAR	100	7089	1772
FEEDING	100	10,090	2522

B.2 Experimental results

In Tab. 4(a), we compare the performance of the various algorithms with rank $K = 30$. The regularization strength α is set to its optimal value for the LR algorithm, and $|\mathcal{A}^-|/|\mathcal{A}^+| = 1/2$. This allows us to compare the LR algorithm to its "augmented" negative versions without hyperparameter tuning. As PROD performs much worse than LR, it is not included in further experiments.

We evaluate the precision at k as

$$p@k = \frac{1}{|\mathcal{T}|}\sum_{A\in\mathcal{T}} \frac{1}{|A|}\sum_{i\in A} \mathbb{1}[\text{rank}(i \mid A\backslash\{i\}) \leq k].$$

Table 4: MPR, p@k, and AUC values for LR, and baseline improvement over LR for other methods. Positive values indicate the algorithm performs better than LR, and bold values indicate improvement over LR that lies outside the standard deviation. Experiments were run 5 times, with $|\mathcal{A}^+|/|\mathcal{A}^-| = \frac{1}{2}$; α is set to its optimal LR value.

		Improvement over LR		
Metric	LR	DYN	EXP	NCE
MPR	70.50	**0.92 ± 0.56**	**0.68 ± 0.62**	**0.86 ± 0.55**
p@1	9.96	0.67 ± 0.75	0.58 ± 0.76	0.20 ± 1.75
p@5	25.36	**1.04 ± 0.82**	**0.78 ± 0.67**	0.67 ± 1.09
p@10	36.50	**1.39 ± 0.85**	**1.13 ± 0.79**	0.97 ± 1.18
p@20	51.22	**1.38 ± 0.97**	**1.28 ± 1.11**	**1.35 ± 1.20**
AUC	0.630	**0.027 ± 0.017**	**0.026 ± 0.016**	0.009 ± 0.017

Compared to traditional SGA methods, algorithms that use inferred negatives perform (PROD excepted) better across all metrics and datasets. DYN and EXP provide consistent improvements compared to the other methods, whereas NCE shows a higher variance and slightly worse performance. Improvements observed using DYN and EXP are larger than the loss in performance due to going from full-rank to low-rank kernels reported in [Gartrell et al., 2017].

Finally, we also compared all methods when tuning both the regularization α and the negative to positive ratio $|\mathcal{A}^-|/|\mathcal{A}^+|$, but did not see any significant improvements. As this suggests there is no need to do additional hyperparameter tuning when using CE, we fix $\frac{|\mathcal{A}^-|}{|\mathcal{A}^+|} = \frac{1}{2}$ for all experiments.