Text-Deconfounded Factorization for Selection-Biased Ratings

Jianfeng Ma

STCS 6701: Probabilistic Models and Machine Learning

1. Problem Formulation

Let $Y \in \mathbb{R}^{N \times M}$ be the rating matrix and $A_{ui}^{\text{train}} \in \{0,1\}^{N \times M}$ be the exposure matrix. Ratings are **Missing-Not-At-Random (MNAR)**:

$$P(A_{ui}^{\text{train}} = 1 \mid Y_{ui}) \neq P(A_{ui}^{\text{train}} = 1)$$

Latent confounders U affect both exposure A and outcome Y. Standard MF minimizes error only on observed entries $\mathcal{O}$, leading to biased estimators.

2. The TDFM Model

We augment Matrix Factorization with two causal surrogates.

The Predictor:

$$\hat{y}_{ui} = 1 + 4 \cdot \sigma\left(\mathbf{u}_u^\top \mathbf{v}_i + b_u + b_i + \beta_z(z_{ui}) + \beta_t(\theta_{ui})\right)$$

Surrogate 1: Exposure surrogate (Poisson factorization) We fit a Poisson factorization exposure model on training observations only:

$$A_{ui}^{\text{train}} \sim \text{Poisson}(\lambda_{ui}), \qquad \lambda_{ui} = s_u^\top t_i, \quad s_u, t_i \in \mathbb{R}_+^{K_{\text{PF}}}.$$

We define the exposure surrogate as $z_{ui} := \hat{\lambda}_{ui}$, and use a linear exposure adjustment: $\beta_z(z_{ui}) = \gamma\, z_{ui}$.

Surrogate 2: Text-Topic Proxy From review text w_{ui}, we infer topic proportions $\theta_{ui} \in \Delta^{K-1}$:

$$x_{ui} = \text{BoW}(w_{ui}), \qquad h_{ui} = \text{ReLU}(\mathbf{W}_1 x_{ui}), \qquad \theta_{ui} = \text{softmax}(\mathbf{W}_2 h_{ui}).$$

We use a linear topic adjustment: $\beta_t(\theta_{ui}) = \alpha^\top \theta_{ui}$.

3. Probabilistic Graphical Model (PGM)

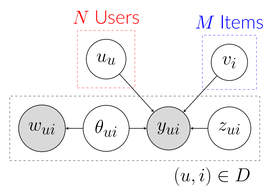


Figure 1: Causal Graph. Surrogates z_{ui} and θ block back-door paths.

Generative Process:

1. Draw latents $u_u, v_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.
2. For (u,i), if observed $(A_{ui} = 1)$:
   - Draw text $w_{ui} \sim \text{Multinomial}(\text{Topic}(v_i))$.
   - Form proxies:

$$z_{ui} = \phi(u,i)$$

$$\theta_{ui} \leftarrow \text{Encoder}(w_{ui})$$

   - Draw rating $y_{ui} \sim \mathcal{N}(\mu_{TDFM}, \epsilon^2)$.

Model Notation:

Sym.	Space	Description
N, M	$\mathbb{N}$	Users (2021) / Items (100)
K, d	$\mathbb{N}$	Topics (10) / Latents (8)
z_{ui}	$\mathbb{R}^1$	Exposure Surrogate
θ_{ui}	Δ^{K-1}	Topic Proportions
$\beta_z(z_{ui})$	$\mathbb{R}$	Exposure adjustment ($= \gamma z_{ui}$)

Theoretical Identification: To recover unbiased preferences from MNAR data, we require the *Proxy Ignorability* assumption. We posit that the learned surrogates form a valid adjustment set $S = \{z_{ui}, \theta_{ui}\}$ that satisfies the **Back-Door Criterion**:

$$\text{(Proxy ignorability)} \quad Y_{ui}(a) \perp A_{ui} \mid z_{ui}, \theta_{ui}$$

- **Exposure Proxy (z):** Captures popularity-driven bias (e.g., "everyone sees this item").
- **Text Proxy (θ):** Captures content-driven bias (e.g., "I only review spicy food").

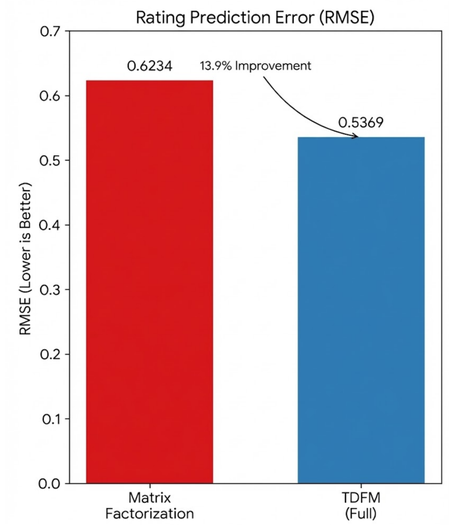
By conditioning on S, we block the confounding path $U \to A \to Y$, ensuring that the remaining correlation between user/item factors solely reflects true preference.

4. Implementation Details

Hyperparameters for the final TDFM training run on the dense subset.

Parameter	Value
Latent Dimension (d)	8
Components (K)	10
Exposure Adjustment	$\beta_z(z_{ui}) = \gamma z_{ui}$
Optimizer	Adam ($lr = 1e-2$)
Batch Size	64
Reg. (weight decay)	$1e-4$

5. Rating Prediction (RMSE)



Key Result: TDFM reduces RMSE by **13.9%** over standard MF.

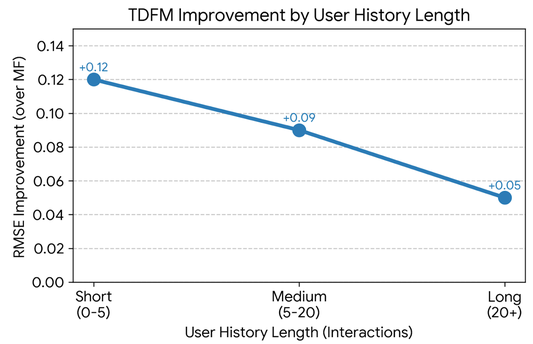
Statistics:

- $\Delta\text{RMSE}: 0.0865$ (95% CI $[0.0611, 0.1138]$)
- Test: Paired bootstrap (by user, $B = 2000$)

Interpretation: The causal surrogates successfully capture and remove bias, allowing the model to learn true underlying preferences.

6. Descriptive Slice (Cold Start)

In a user-history slice, TDFM shows larger RMSE gains for sparse-history users.



7. Training Algorithm

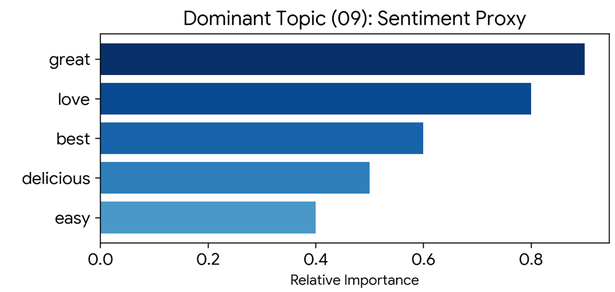
We optimize MF factors and linear surrogate adjustments (topic + exposure) end-to-end.

Algorithm 1 TDFM Training (linear surrogates)

Input: observed ratings Y, exposure indicator A, review text W
Pre-compute: exposure surrogate $z_{ui} = \phi(u,i)$ (e.g., log-degrees)
Pre-train: topic encoder on text to obtain topic proportions $\theta_{ui} = \text{Encoder}(w_{ui})$
Initialize: user/item factors U, V, topic weights α, exposure weight γ
for epoch $= 1$ to E **do**
  for batch $(u,i) \in \mathcal{O}$ **do**
    Compute score $s_{ui} = u_u^\top v_i + \alpha^\top \theta_{ui} + \gamma z_{ui}$
    Predict $\hat{y}_{ui} = 1 + 4\sigma(s_{ui})$
    $\mathcal{L} \leftarrow (y_{ui} - \hat{y}_{ui})^2 + \lambda\left(\|u_u\|_2^2 + \|v_i\|_2^2 + \|\alpha\|_2^2 + \gamma^2\right)$
    Backpropagate and update U, V, α, γ (Adam)
  end for
end for

8. Diagnostic: Topic Collapse

Why does text help RMSE but hurt ranking?



- **Observation:** Topic 09 captures 80% of probability mass.
- **Insight:** The model learned a **Sentiment Proxy** (e.g., "Great", "Love") rather than a semantic confounder.
- **Consequence:** This acts as a global intercept, shifting predicted ratings up (helping RMSE) but failing to distinguish relative order.

9. Conclusion & Future Work

Summary of Contributions:

- **Causal Adjustment:** Explicitly adjust for exposure and text proxies (z_{ui}, θ_{ui}) to mitigate MNAR selection bias.
- **The Text Trade-off:** While text reduces intensity error, standard topic models suffer from *sentiment collapse*, acting as a proxy for rating magnitude rather than content preference.

Future Directions:

- **Adversarial Unlearning:** Implement a gradient reversal layer to force learned topics to be orthogonal to sentiment labels.
- **Semantic Disentanglement:** Replace BoW with pre-trained sentence embeddings (e.g., S-BERT) to capture semantic context independent of valence.
- **Large-Scale Validation:** Evaluate robustness on 1M+ user datasets (e.g., Amazon Books).