Ordinal Non-negative Matrix Factorization for Recommendation

International Conference on Machine Learning



Olivier Gouvert[1] Thomas Oberlin[2] Cédric Févotte[1]

[1]IRIT, Université de Toulouse, CNRS, France
[2]ISAE-SUPAERO, Université de Toulouse, France

Collaborative Filtering (CF)

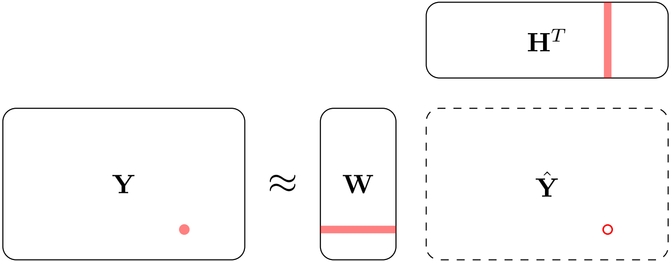
- Based only on the feedbacks of users on items
- $\mathbf{Y}$: feedback matrix, of size $U \times I$
  y_{ui}: feedback of a user $u \in \{1, \ldots, U\}$ on an item $i \in \{1, \ldots, I\}$

Collaborative Filtering (CF)

- Based only on the feedbacks of users on items
- $\mathbf{Y}$: feedback matrix, of size $U \times I$
  y_{ui}: feedback of a user $u \in \{1, \ldots, U\}$ on an item $i \in \{1, \ldots, I\}$
- **Ordinal data**: nominal data which exhibit a natural ordering [Stevens et al., 1946]:
  - Explicit feedbacks: bad $\prec$ average $\prec$ good $\prec$ excellent
  - Implicit feedbacks: quantized play counts
  - ... without loss of generality $y_{ui} \in \{0, \ldots, V\}$

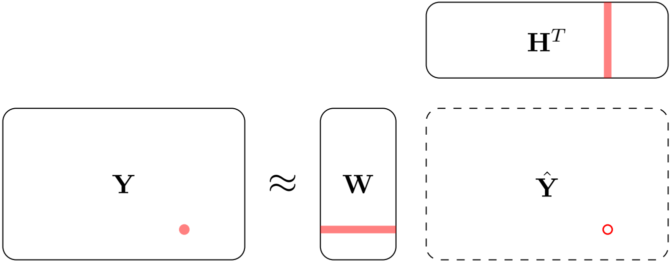
Non-negative Matrix Factorization (NMF)

- Approximation: $\mathbf{Y} \approx \mathbf{W}\mathbf{H}^T$ [Lee and Seung, 1999]
  - $\mathbf{W} \geq 0$ of size $U \times K$: preferences of the users
  - $\mathbf{H} \geq 0$ of size $I \times K$: attributes of the items



Non-negative Matrix Factorization (NMF)

- Approximation: $\mathbf{Y} \approx \mathbf{W}\mathbf{H}^T$ [Lee and Seung, 1999]
  - $\mathbf{W} \geq 0$ of size $U \times K$: preferences of the users
  - $\mathbf{H} \geq 0$ of size $I \times K$: attributes of the items



- How to process ordinal data?

How to process ordinal data?

- **Threshold models:**
  - Quantization of a continuous latent variable
  - Some examples: [Chu and Ghahramani, 2005, Paquet et al., 2012, Hernandez-Lobato et al., 2014]

How to process ordinal data?

- **Threshold models:**
  - Quantization of a continuous latent variable
  - Some examples: [Chu and Ghahramani, 2005, Paquet et al., 2012, Hernandez-Lobato et al., 2014]

- **Contributions:**
  - NMF for ordinal data (OrdNMF)
    - Non-negative constraints
    - Multiplicative noise
    - Link with Poisson factoriaztion (PF) [Gopalan et al., 2015]
  - Efficient variational algorithm
    - Augmentation trick
    - Scales with the number of non-zero values
  - Excellent flexibility of OrdNMF

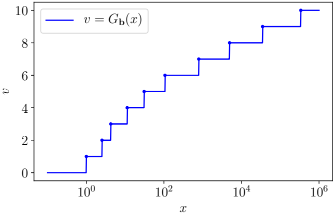
Ordinal Non-negative Matrix Factorization (OrdNMF)

- Approximation: $\mathbf{Y} \approx G_{\mathbf{b}}(\mathbf{W}\mathbf{H}^T)$
  - $\mathbf{Y}$ ordinal matrix
  - $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$

- **Quantization of the non-negative numbers**

$$G_{\mathbf{b}} : \begin{array}{ccc} \mathbb{R}_+ & \rightarrow & \{0, \dots, V\} \\ x & \mapsto & v \text{ such that } x \in [b_{v-1}, b_v) \end{array}$$

where $\mathbf{b}$ is an increasing sequence of thresholds



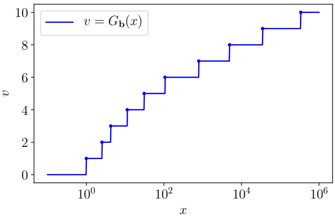
Ordinal Non-negative Matrix Factorization (OrdNMF)

- Approximation: $\mathbf{Y} \approx G_{\mathbf{b}}(\mathbf{W}\mathbf{H}^T)$
  - $\mathbf{Y}$ ordinal matrix
  - $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$

- **Quantization of the non-negative numbers**

$$G_{\mathbf{b}} : \begin{array}{ccl} \mathbb{R}_+ & \to & \{0, \dots, V\} \\ x & \mapsto & v \text{ such that } x \in [b_{v-1}, b_v) \end{array}$$

where $\mathbf{b}$ is an increasing sequence of thresholds

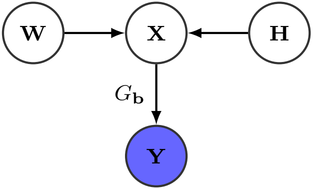


- **Goal:** joint estimation of $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{b}$

Ordinal Non-negative Matrix Factorization (OrdNMF)

- Generative model:

$$x_{ui} = [\mathbf{W}\mathbf{H}^T]_{ui} \cdot \varepsilon_{ui}$$
$$y_{ui} = G_{\mathbf{b}}(x_{ui})$$

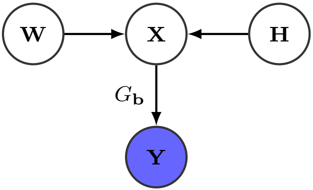


- **Multiplicative noise:** ε non-negative random variable with c.d.f. F_ε

Ordinal Non-negative Matrix Factorization (OrdNMF)

- Generative model:

$$x_{ui} = [\mathbf{W}\mathbf{H}^T]_{ui} \cdot \varepsilon_{ui}$$
$$y_{ui} = G_{\mathbf{b}}(x_{ui})$$



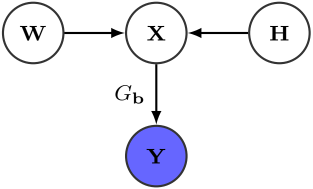
- **Multiplicative noise:** ε non-negative random variable with c.d.f. F_ε
- Cumulative distribution function:

$$\begin{aligned}\mathbb{P}[y_{ui} \leq v|\mathbf{W},\mathbf{H}] &= \mathbb{P}\left[G_{\mathbf{b}}(x_{ui}) \leq v|\mathbf{W},\mathbf{H}\right] \\ &= \mathbb{P}\left[[\mathbf{W}\mathbf{H}^T]_{ui} \cdot \varepsilon_{ui} < b_v\right] \\ &= \mathbb{P}\left[\varepsilon_{ui} < \frac{b_v}{[\mathbf{W}\mathbf{H}^T]_{ui}}\right] \\ &= F_\varepsilon\left(\frac{b_v}{[\mathbf{W}\mathbf{H}^T]_{ui}}\right)\end{aligned}$$

Inverse-Gamma OrdNMF

- Generative model:

$$x_{ui} = [\mathbf{W}\mathbf{H}^T]_{ui} \cdot \varepsilon_{ui}$$
$$y_{ui} = G_{\mathbf{b}}(x_{ui})$$

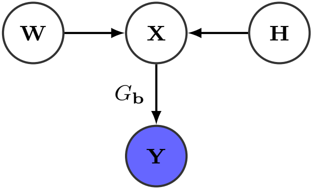


- **Inverse-gamma noise:** $\varepsilon_{ui} \sim \mathrm{IG}(1, 1)$

Inverse-Gamma OrdNMF

- Generative model:

$$x_{ui} = [\mathbf{W}\mathbf{H}^T]_{ui} \cdot \varepsilon_{ui}$$
$$y_{ui} = G_{\mathbf{b}}(x_{ui})$$



- **Inverse-gamma noise:** $\varepsilon_{ui} \sim \mathrm{IG}(1,1)$
- Cumulative distribution function:

$$\mathbb{P}[y_{ui} \leq v | \mathbf{W}, \mathbf{H}] = e^{-[\mathbf{W}\mathbf{H}^T]_{ui} b_v^{-1}}$$
$$\text{or } \mathbb{P}[y_{ui} > v | \mathbf{W}, \mathbf{H}] = 1 - e^{-[\mathbf{W}\mathbf{H}^T]_{ui} b_v^{-1}}$$

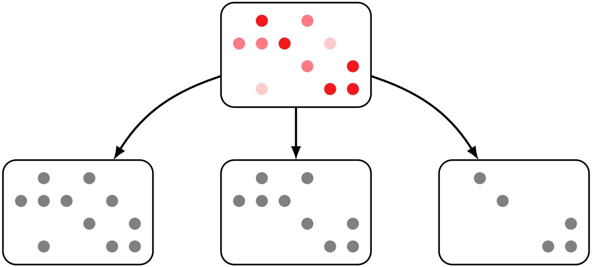
Interpretation

- V dependent Bernoulli models:

$$\{y_{ui} > v\} \sim \text{Bern}\left(1 - e^{-[\mathbf{WH}^T]_{ui} b_v{}^{-1}}\right), \quad v \in \{0, \dots, V-1\}$$

  - $V = 1$: Bernoulli-Poisson factorization (BePoF) [Acharya et al., 2015]
  - ... Poisson factorization (PF) [Gopalan et al., 2015] applied on binary data

$\mathbf{Y}$ with $V = 3$ (4 classes)



$\mathbf{Y} > 0$ $\mathbf{Y} > 1$ $\mathbf{Y} > 2$

Bayesian Inference

- Bayesian inference:
  - A priori: $w_{uk} \sim \mathrm{Gamma}(\alpha^W, \beta_u^W)$ and $h_{ik} \sim \mathrm{Gamma}(\alpha^H, \beta_i^H)$
  - Variational inference (VI): $p(\mathbf{W}, \mathbf{H}|\mathbf{Y}) \approx q(\mathbf{W})q(\mathbf{H})$

Bayesian Inference

- Bayesian inference:
  - A priori: $w_{uk} \sim \text{Gamma}(\alpha^W, \beta_u^W)$ and $h_{ik} \sim \text{Gamma}(\alpha^H, \beta_i^H)$
  - Variational inference (VI): $p(\mathbf{W}, \mathbf{H}|\mathbf{Y}) \approx q(\mathbf{W})q(\mathbf{H})$
- Log-likelihood, with $\Delta_v = b_{v-1}^{-1} - b_v^{-1}$:

$$\log \mathbb{P}[y_{ui} = v|\mathbf{W}, \mathbf{H}] = \begin{cases} -[\mathbf{W}\mathbf{H}^T]_{ui}\ b_0^{-1}, & \text{if } v = 0 \\ -[\mathbf{W}\mathbf{H}^T]_{ui}\ b_v^{-1} + \log(1 - e^{-[\mathbf{W}\mathbf{H}^T]_{ui}\Delta_v}), & \text{if } v > 0 \end{cases}$$

Non-conjugate model

Model Augmentation

- **Trick:** model augmentation similar to [Acharya et al., 2015]

$$n_{ui}|y_{ui}, \mathbf{W}, \mathbf{H} \sim \begin{cases} \delta_0, & \text{if } y_{ui} = 0 \\ \mathrm{ZTP}([\mathbf{W}\mathbf{H}^T]_{ui}\Delta_{y_{ui}}), & \text{if } y_{ui} > 0 \end{cases}$$

  - Joint likelihood: generalized Kullback-Leibler divergence
  - Scales with the number of non-zero in $\mathbf{Y}$

Model Augmentation

- **Trick:** model augmentation similar to [Acharya et al., 2015]

$$n_{ui}|y_{ui}, \mathbf{W}, \mathbf{H} \sim \begin{cases} \delta_0, & \text{if } y_{ui} = 0 \\ \text{ZTP}([\mathbf{W}\mathbf{H}^T]_{ui}\Delta_{y_{ui}}), & \text{if } y_{ui} > 0 \end{cases}$$

  - Joint likelihood: generalized Kullback-Leibler divergence
  - Scales with the number of non-zero in $\mathbf{Y}$

- **Threshold optimization:** working on the decrement sequence $\boldsymbol{\Delta}$ (defined as $\Delta_v = b_{v-1}^{-1} - b_v^{-1}$) rather than on the threshold sequence $\mathbf{b}$
  - Very simple update rules for $\mathbf{b}$

Experimental Results

- MovieLens dataset
  - Ratings of users on movies on a scale from 1 to 10
  - Class 0: absence of a rating
- Splitting of $\mathbf{Y}$:
  - $\mathbf{Y}^{\text{train}}$: 80% of the non-zero values
  - $\mathbf{Y}^{\text{test}}$: remaining 20%

Prediction Results

- Normalized discounted cumulative gain (NDCG)
  - Ranking metric
  - Relevance: $\mathrm{rel}(u,i) = \mathbb{1}[y_{ui}^{\text{test}} \geq s]$

Table: Recommendation performance. R: raw data. B: binary data

			NDCG @100 with threshold s				
Model	Data	K	$s=1$	$s=4$	$s=6$	$s=8$	$s=10$
OrdNMF	R	150	**0.444**	**0.444**	**0.439**	**0.414**	0.353
BePoF	B (≥ 1)	50	0.433	0.430	0.421	0.383	0.310
PF	B (≥ 1)	100	0.431	0.428	0.418	0.380	0.306
BePoF	B (≥ 8)	50	0.389	0.393	0.399	0.408	**0.369**
PF	B (≥ 8)	150	0.386	0.389	0.395	0.403	0.365

Posterior Predictive Check (PPC)

- Generating new data based on the posterior predictive distribution $p(\mathbf{Y}^*, \mathbf{W}, \mathbf{H}|\mathbf{Y}) \approx p(\mathbf{Y}^*|\mathbf{W}, \mathbf{H})q(\mathbf{W})q(\mathbf{H})$

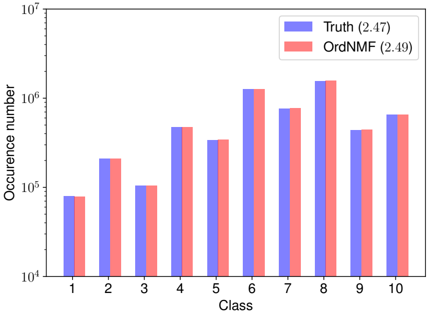


Figure: PPC of the distribution of the classes in the MovieLens dataset

Conclusion

- **Take-home message:**
  - NMF framework to process ordinal data
  - Natural extension of BePoF
  - Efficient variational algorithm - scales with non-zero values
  - Flexibility of OrdNMF

- **More information:**
  - GitHub: `https://github.com/Oligou/OrdNMF`
  - Contact: oliviergouvert@gmail.com

References I

[Acharya et al., 2015] Acharya, A., Ghosh, J., and Zhou, M. (2015).
Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices.
In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Chu and Ghahramani, 2005] Chu, W. and Ghahramani, Z. (2005).
Gaussian processes for ordinal regression.
The Journal of Machine Learning Research, 6(Jul):1019–1041.

[Gopalan et al., 2015] Gopalan, P., Hofman, J. M., and Blei, D. M. (2015).
Scalable Recommendation with Hierarchical Poisson Factorization.
In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 326–335.

[Hernandez-Lobato et al., 2014] Hernandez-Lobato, J. M., Houlsby, N., and Ghahramani, Z. (2014).
Probabilistic Matrix Factorization with Non-random Missing Data.
In *Proc. International Conference on Machine Learning (ICML)*, pages 1512–1520.

References II

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999).
Learning the Parts of Objects by Non-negative Matrix Factorization.
Nature, 401(6755):788–791.

[Paquet et al., 2012] Paquet, U., Thomson, B., and Winther, O. (2012).
A hierarchical model for ordinal matrix factorization.
Statistics and Computing, 22(4):945–957.

[Stevens et al., 1946] Stevens, S. S. et al. (1946).
On the theory of scales of measurement.
Science.