UNIVERSITY OF CAMBRIDGE

Probabilistic Matrix Factorization with Non-random Missing Data

José Miguel Hernández-Lobato, Neil Houlsby and Zoubin Ghahramani
Engineering Department, University of Cambridge, United Kingdom.

ICML | Beijing
International conference on machine learning, 2014

1. Introduction

Motivation: There is empirical evidence that real-world **rating data** is **missing not at random** (MNAR): the random process that selects the observed data depends on the value of the unobserved ratings.

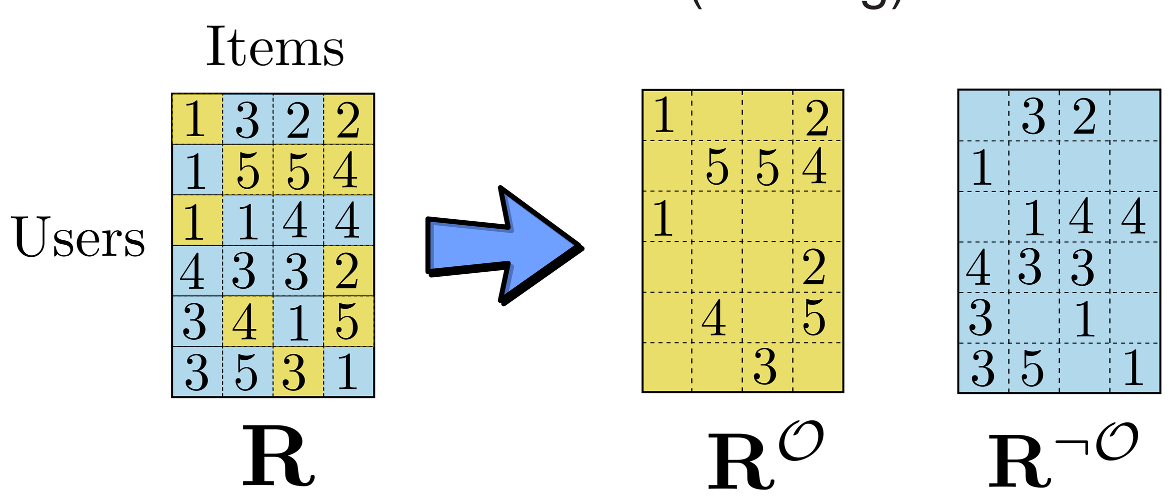
Problem: **Probabilistic matrix factorization** (MF) models have state-of-the-art predictive performance. However, they often assume **missing at random** (MAR) rating data and ignore dependencies.

Solution: first practical implementation of a probabilistic MF model for ordinal MNAR rating data (**MF-MNAR**). Dependencies captured by combining a complete data model (**CDM**) with a missing data model (**MDM**). Scalability achieved by using **stochastic inference** methods.

2. Ordinal Rating Data and Matrices

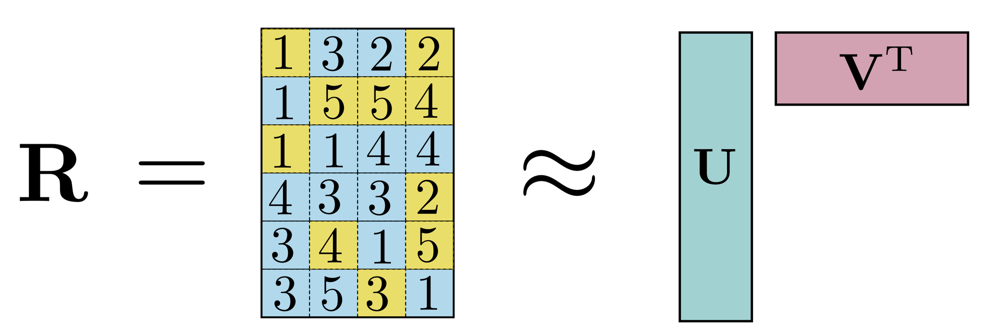
Large amounts of **ordinal rating data** produced on the Internet:

Encoded as a **matrix** $\mathbf{R}$ with observed (missing) entries $\mathbf{R}^{\mathcal{O}}$ ($\mathbf{R}^{\neg\mathcal{O}}$).



3. Matrix Factorization Models and the MAR Assumption

We expect $\mathbf{R}$ to be well approximated by the **low rank** matrix $\mathbf{UV}^{\text{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ have a small number of columns:



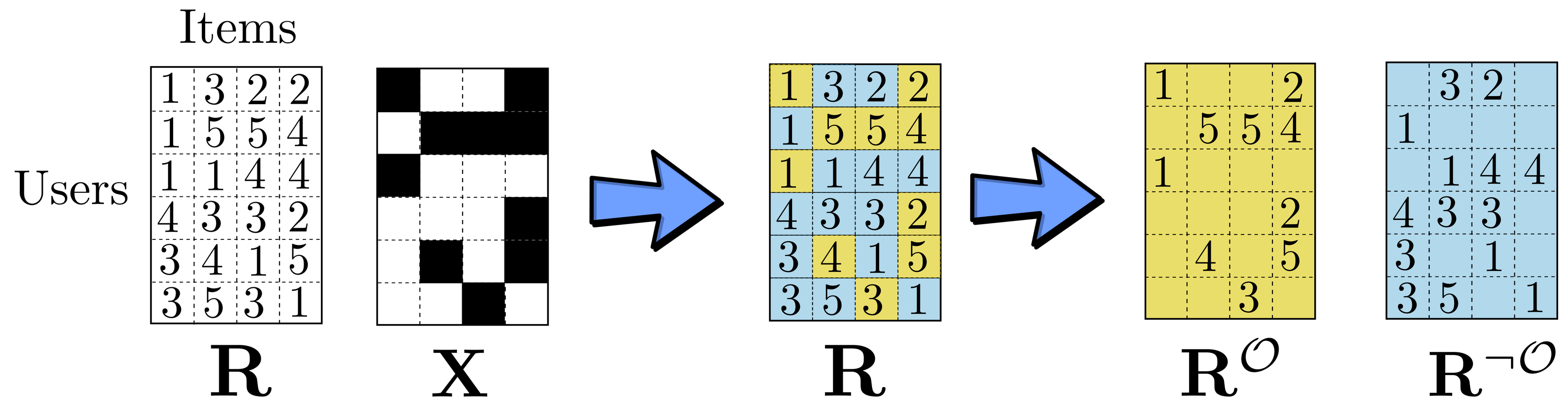
We can **marginalize out** the missing entires in the likelihood to obtain

$$p(\mathbf{R}^{\mathcal{O}}|\mathbf{U},\mathbf{V}) = \sum_{\mathbf{R}^{\neg\mathcal{O}}} p(\mathbf{R}^{\mathcal{O}},\mathbf{R}^{\neg\mathcal{O}}|\mathbf{U},\mathbf{V}) = \sum_{\mathbf{R}^{\neg\mathcal{O}}} \prod_{i=1}^{n}\prod_{j=1}^{d} p(r_{i,j}|\mathbf{u}_i\mathbf{v}_j^{\text{T}}) = \prod_{(i,j)\in\mathcal{O}} p(r_{i,j}|\mathbf{u}_i\mathbf{v}_j^{\text{T}})\,.$$

This introduces independence assumptions! MAR assumption!

3. Probabilistic Treatment of Missing Data

We use a binary matrix $\mathbf{X}$ that splits $\mathbf{R}$ into $\mathbf{R}^{\mathcal{O}}$ and $\mathbf{R}^{\neg\mathcal{O}}$.



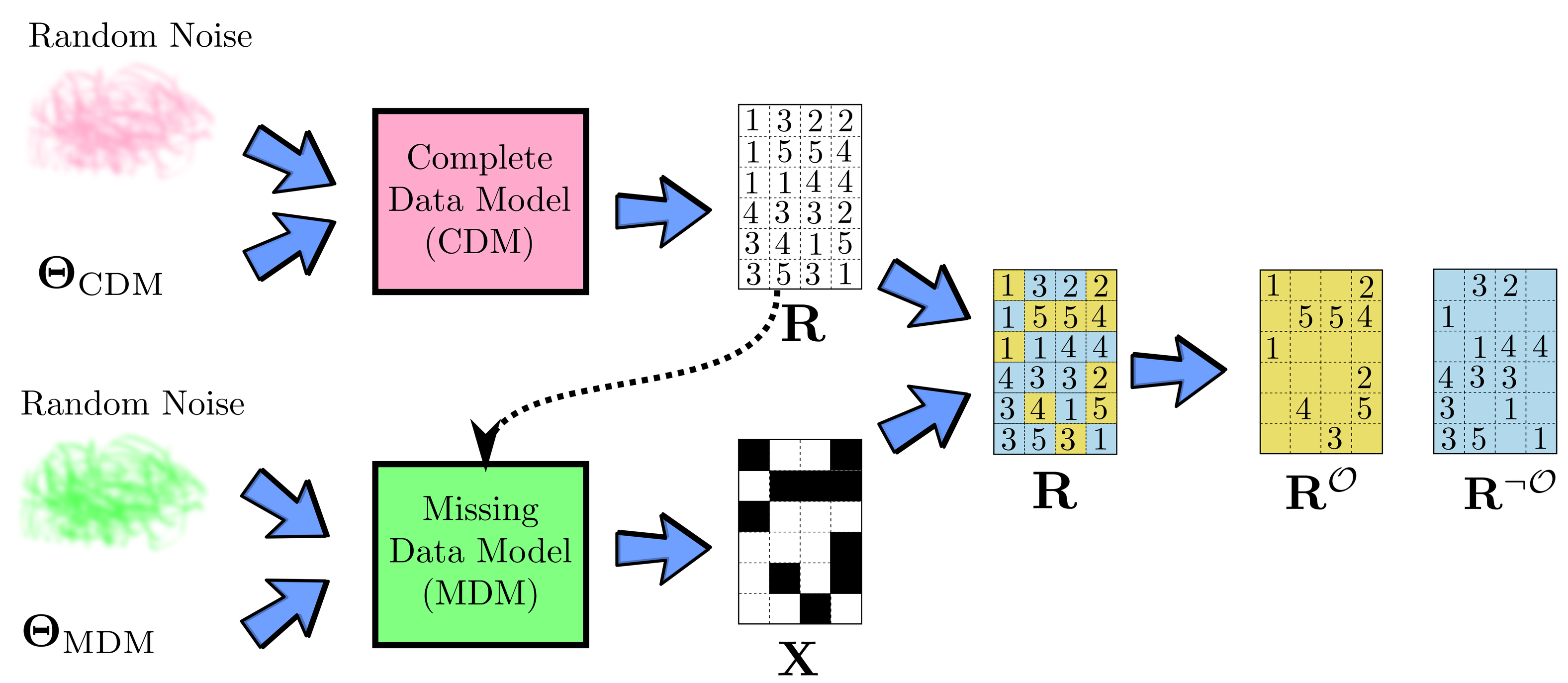
We use a **joint probabilistic model** for $\mathbf{R}$ and $\mathbf{X}$ formed by

- A complete data model (CDM) for $\mathbf{R}$: $p(\mathbf{R}|\Theta_{\text{CDM}})$.
- A missing data model (MDM) for $\mathbf{X}$ given $\mathbf{R}$: $p(\mathbf{X}|\mathbf{R},\Theta_{\text{MDM}})$.

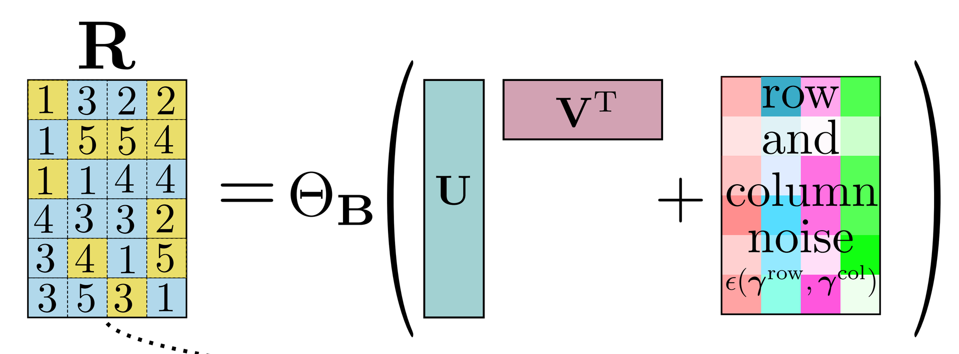
4. The Joint Model for the Data

The joint likelihood for Θ_{CDM} and Θ_{MDM} given $\mathbf{X}$ and $\mathbf{R}$ is

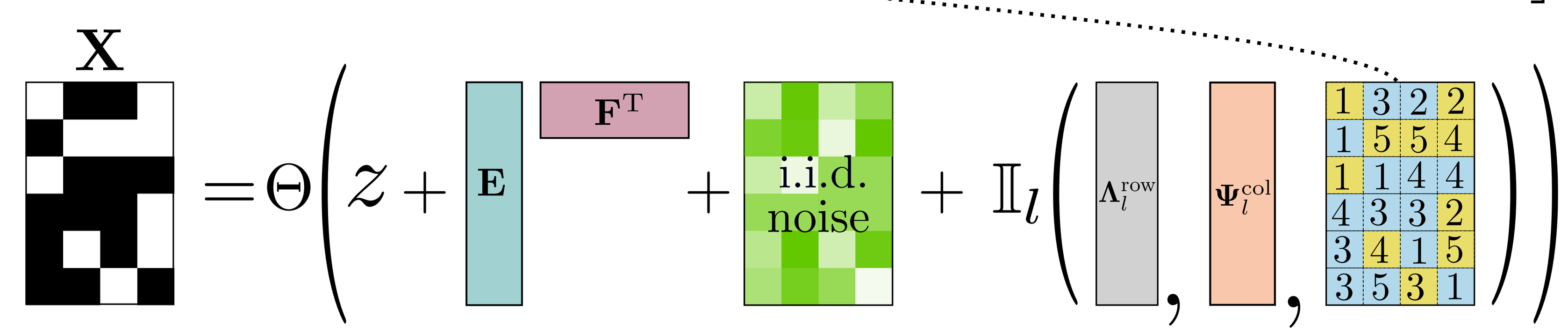
$$p(\mathbf{X},\mathbf{R}|\Theta_{\text{CDM}},\Theta_{\text{MDM}}) = p(\mathbf{X}|\mathbf{R},\Theta_{\text{MDM}})\,p(\mathbf{R}|\Theta_{\text{CDM}})\,.$$



6. A Diagram with our Implementation for the Joint Model



$$p(\mathbf{X}|\mathbf{E},\mathbf{F},z,\boldsymbol{\Lambda}^{\text{row}},\boldsymbol{\Psi}^{\text{col}},\mathbf{R}) = \left[\prod_{i=1}^{n}\prod_{j=1}^{d}\sigma\{(2x_{i,j}-1)(\mathbf{e}_i\mathbf{f}_j^{\text{T}} + z + \sum_{l=1}^{L}(\lambda_{i,l}^{\text{row}}+\psi_{j,l}^{\text{col}})\mathbf{I}[r_{i,j}=l])\}\right]$$



7. Inference Algorithm and Predictive Distribution

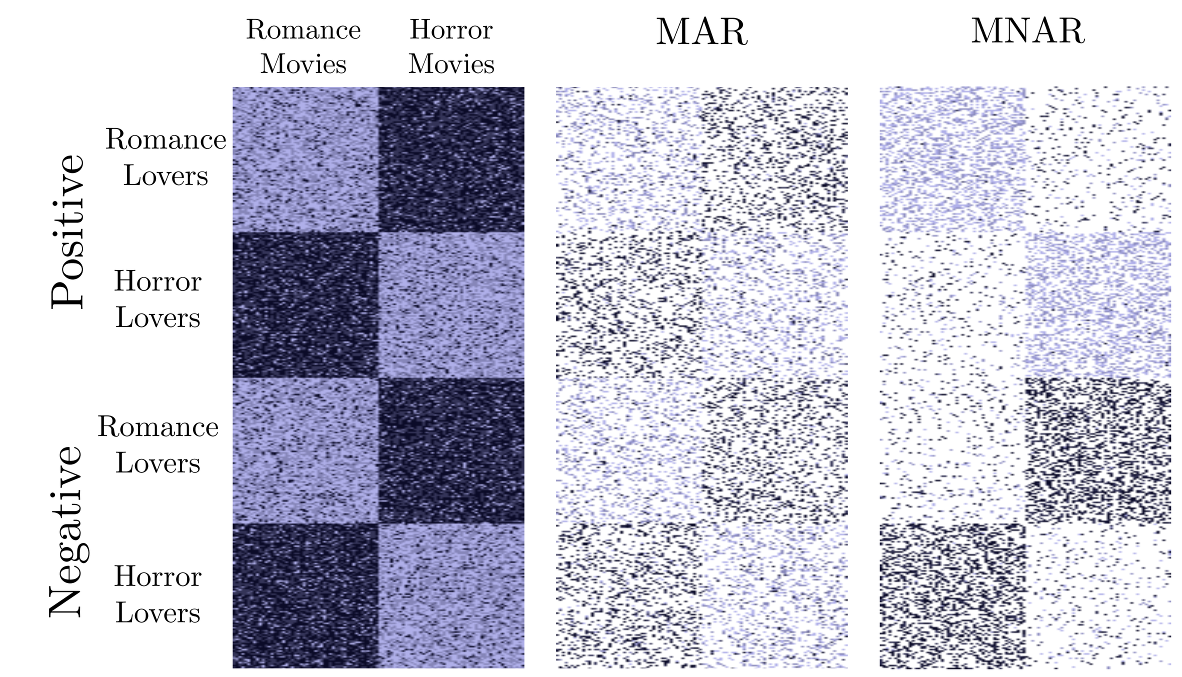
We adjust the **posterior approximation** $\mathcal{Q}$ using

> **Input:** Rating dataset $\mathcal{D}$.
> Adjust $\mathcal{Q}$ using EP and VB on CDM ignoring MDM.
> Adjust $\mathcal{Q}$ using SVI on MDM ignoring CDM.
> **for** $t = 1$ **to** T **do**
>   Adjust $\mathcal{Q}$ using SVI on MDM.
>   Adjust $\mathcal{Q}$ using EP-SVI on CDM.
> **end for**
> **Output:** Posterior approximation $\mathcal{Q}$.

Given $\mathcal{Q}$, the **predictive distribution** for $r_{i,j}$ is a function of $x_{i,j}$:

$$p(r_{i,j}=l|\mathbf{R}^{\mathcal{O}},\mathbf{X}) \approx \tilde{p}_{i,j,l}^{\text{JM}}(x_{i,j}) \propto \tilde{p}_{i,j,l}^{\text{CDM}}\tilde{p}_{i,j,l}^{\text{MDM}}(x_{i,j})\,,$$

8. The Synthetic SRH Dataset



9. Predictive Performance on $\mathbf{R}^{\neg\mathcal{O}}$ and X

Results on $\mathbf{R}^{\neg\mathcal{O}}$:

MF-MNAR: CDM & MDM.

MF-MAR: CDM only & MAR assumption.

Table : Average Log-likelihood on the Standard Test Sets.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
ML100K	**-1.181**	-1.186	-1.471	-1.463	-1.425	-1.218	-1.468
ML1M	**-1.121**	-1.125	-1.308	-1.436	-1.380	-1.162	-1.456
MTweet	**-0.941**	-0.946	-1.105	-1.245	-1.141	-0.997	-1.235
NIPS	**-0.937**	-0.956	-1.204	-1.170	-1.167	-0.995	-1.329
Yahoo	**-1.172**	-1.204	-1.278	-1.399	-1.304	-1.218	-1.551
SMF-MNAR	**-0.902**	-0.937	-1.447	-1.336	-1.326	-1.000	-1.331
SMF-MAR	-0.425	**-0.417**	-1.327	-1.238	-1.235	-0.510	-1.198
SRH-MNAR	-1.055	-1.067	-0.987	**-0.962**	-0.963	-1.143	-1.392
SRH-MAR	-1.317	-1.287	-1.272	**-1.265**	-1.266	-1.318	-1.498

Results on X:

MF-MNAR: CDM & MDM.

MDM: MDM only.

Table : Average Recall on the Standard Test Sets.

Dataset	MF MNAR	MDM	CTPv MNAR	Logitvd MNAR	Freq
ML100K	**0.299**	0.295	0.093	0.130	0.119
ML1M	0.204	**0.204**	0.041	0.068	0.077
MTweet	**0.203**	0.199	0.127	0.142	0.143
NIPS	**0.309**	0.309	0.011	0.009	0.013
Yahoo	**0.285**	0.283	0.145	0.198	0.182
SMF-MNAR	**0.300**	0.280	0.038	0.052	0.051
SMF-MAR	0.438	**0.445**	0.038	0.051	0.047
SRH-MNAR	**0.246**	0.245	0.209	0.157	0.121
SRH-MAR	0.113	0.115	0.134	**0.143**	0.131