Recommendation on Data Missing Not at Random

A Doubly Robust Joint Learning Approach

Xiaojie Wang[1], Rui Zhang[1], **Yu Sun**[2], and Jianzhong Qi[1]

[1]University of Melbourne, [2]Twitter

Rating Matrix

	Item 1	Item 2	Item 3	...	Item M
User 1	4			...	
User 2			2	...	
User 3		5		...	5
...	...	...	...	...	...
User N			2	...	1

Rating Prediction

	Item 1	Item 2	Item 3	...	Item M
User 1	4.5	2.3	3.5	...	1.8
User 2	6.7	3.9	2.9	...	3.8
User 3	2.3	4.8	1.1	...	5.2
...	...	...	...	...	...
User N	2.6	3.5	1.8	...	0.7

Prediction Error

	Item 1	Item 2	Item 3	...	Item M
User 1	4.5 - 4 = 0.5			...	
User 2			2.9 - 2 = 0.9	...	
User 3		5 - 4.8 = 0.2		...	5.2 - 5 = 0.2
...	...	...	...	...	...
User N			2 - 1.8 = 0.2	...	1 - 0.7 = 0.3

Prediction Error

	Item 1	Item 2	Item 3	...	Item M
User 1	4.5 - 4 = 0.5	2.3	3.5	...	1.8
User 2	6.7	3.9	2.9 - 2 = 0.9	...	3.8
User 3	2.3	5 - 4.8 = 0.2	1.1	...	5.2 - 5 = 0.2
...	...	...	...	...	...
User N	2.6	3.5	2 - 1.8 = 0.2	...	1 - 0.7 = 0.3

Handling Missing Ratings: Ignore Them

$$\frac{1}{|\mathcal{D}|}\sum_{u,i\in\mathcal{D}}(o_{u,i}e_{u,i})$$

When missing ratings are **missing at random** (**MAR**), the prediction error is unbiased
i.e.,

$$\mathbb{E}_{\mathbf{O}}\left[\frac{1}{|\mathcal{D}|}\sum_{u,i\in\mathcal{D}}(o_{u,i}e_{u,i})\right]=\frac{1}{|\mathcal{D}|}\sum_{u,i\in\mathcal{D}}e_{u,i}$$

	Item 1	Item 2	Item 3	...	Item M
User 1	0.5			...	
User 2			0.9	...	
User 3		0.2		...	0.2
...	...	...	...	...	...
User N			0.2	...	0.3

Missing Ratings: Missing Not at Random

- Missing ratings: **missing not at random** (**MNAR**)
- Rating for an item is missing or not: the **user's rating for that item**
- Producer:
  - Tens of thousands of items, not randomly chosen to present
  - Selection / ranking / filtering process
- User:
  - Normally don't choose items randomly to watch/buy/visit
  - After watching/buying/visiting, don't choose items randomly to rate, either
    - Rate those they have an opinion

Can we **do better** when ratings are MNAR?

Handling Missing Ratings: Error Imputation

$$\frac{1}{|\mathcal{D}|}\sum_{u,i\in\mathcal{D}}(o_{u,i}e_{u,i}+(1-o_{u,i})\hat{e}_{u,i})$$

The imputed errors can be based on heuristics. For example, in an existing work [Steck 2010]:

$$\hat{e}_{u,i}=\omega|\hat{r}_{u,i}-\gamma|$$

	Item 1	Item 2	Item 3	...	Item M
User 1	0.5	2.2	1.0	...	2.7
User 2	2.2	0.6	0.9	...	0.7
User 3	2.2	0.2	3.4	...	0.2
...	...	...	...	...	...
User N	1.9	1.0	0.2	...	0.3

$\omega = 1 \quad \gamma = 4.5$

If the imputed errors are accurate, the prediction error is unbiased

Handling Missing Ratings: Inverse Propensity

$$\frac{1}{|\mathcal{D}|} \sum_{u,i \in \mathcal{D}} \frac{o_{u,i} e_{u,i}}{\hat{p}_{u,i}}$$

where

$$p_{u,i} = P(o_{u,i} = 1 | r_{u,i}, \boldsymbol{x}_{u,i})$$

	Item 1	Item 2	Item 3	...	Item M
User 1	0.5*1.3			...	
User 2			0.9*2.7	...	
User 3		0.2*3.4		...	0.2*1.4
...	...	...	...	...	...
User N			0.2*3.9	...	0.3*1.2

If the estimated propensities are accurate, the prediction error is unbiased

Weakness

- Error imputation based (EIB)
  - **Hard to accurately estimate** the imputed errors
  - it’s almost as hard as predicting the original ratings
- Inverse propensity scoring (IPS)
  - often suffers from the **large variance issue**
  - When estimated propensity is very small, it creates a very large value

Handling Missing Ratings: Proposed Doubly Robust

$$\frac{1}{|\mathcal{D}|}\sum_{u,i\in\mathcal{D}}\left(\frac{o_{u,i}}{\hat{p}_{u,i}}e_{u,i} + (1-\frac{o_{u,i}}{\hat{p}_{u,i}})\hat{e}_{u,i}\right)$$

where

$$p_{u,i} = P(o_{u,i} = 1 | r_{u,i}, \boldsymbol{x}_{u,i})$$

and $\hat{e}_{u,i}$ is the imputed error

	$o_{u,i} = 0$	$o_{u,i} = 1$
$\hat{p}_{u,i}$	$\hat{e}_{u,i}$	$\frac{e_{u,i} - \hat{e}_{u,i}}{\hat{p}_{u,i}} + \hat{e}_{u,i}$
$\hat{p}_{u,i} \to 1$	$\hat{e}_{u,i}$	$e_{u,i}$
$\hat{p}_{u,i} \to 0$	$\hat{e}_{u,i}$	$\approx \hat{e}_{u,i}$ *

* when imputed error is close to the true error

Doubly robust: the prediction error is unbiased when

- **either** the estimated propensities are accurate
- **or** the imputed errors are accurate

Toy Example

True Ratings $\mathbf{R}$

$$\begin{bmatrix} 1 & 1 & 5 \\ 1 & 1 & 5 \end{bmatrix}$$

Predicted Ratings $\hat{\mathbf{R}}$

$$\begin{bmatrix} 3 & 3 & 4 \\ 3 & 3 & 4 \end{bmatrix} \longrightarrow$$

Prediction Errors $\mathbf{E}$

$$\begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \end{bmatrix}$$

Prediction error = 10 / 6

Toy Example

Observation Indicators $\mathbf{O}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Prediction Errors $\mathbf{E}$

$$\begin{bmatrix} 2 & & \\ & & 1 \end{bmatrix}$$

Imputed Errors $\hat{\mathbf{E}}$

$$\begin{bmatrix} 1.5 & 1.5 & 0.5 \\ 1.5 & 1.5 & 0.5 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1.5 & 0.5 \\ 1.5 & 1.5 & 1 \end{bmatrix}$$

Estimated error from EIB is 8 / 6

$$\text{Bias}(\mathcal{E}_{\text{EIB}}) = 0.33$$

Toy Example

Observation Indicators $\mathbf{O}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Prediction Errors $\mathbf{E}$

$$\begin{bmatrix} 2 & & \\ & & 1 \end{bmatrix}$$

Learned Propensities $\hat{\mathbf{P}}$

$$\begin{bmatrix} 0.3 & & \\ & & 0.4 \end{bmatrix}$$

$$\begin{bmatrix} 6.7 & & \\ & & 2.5 \end{bmatrix}$$

Estimated error from IPS is 9.2 / 6

$$\text{Bias}(\mathcal{E}_{\text{IPS}}) = 0.13$$

Toy Example

Observation Indicators $\mathbf{O}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Prediction Errors $\mathbf{E}$

$$\begin{bmatrix} 2 & & \\ & & 1 \end{bmatrix}$$

Imputed Errors $\hat{\mathbf{E}}$

$$\begin{bmatrix} 1.5 & 1.5 & 0.5 \\ 1.5 & 1.5 & 0.5 \end{bmatrix}$$

Learned Propensities $\hat{\mathbf{P}}$

$$\begin{bmatrix} 0.3 & & \\ & & 0.4 \end{bmatrix}$$

$$\begin{bmatrix} 3.17 & 1.5 & 0.5 \\ 1.5 & 1.5 & 1.75 \end{bmatrix}$$

Estimated error from DR is 9.92 / 6

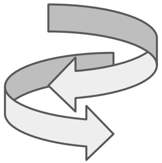
$$\text{Bias}(\mathcal{E}_{\text{DR}}) = 0.01$$

Joint Learning

- Imputed errors are closely related to predicted ratings, e.g., $\hat{e}_{u,i} = \omega|\hat{r}_{u,i} - \gamma|$
  - Accuracy of imputed errors changes when predicted ratings change
  - In turn, changed imputed errors affect rating prediction training
- Joint Learning

Rating prediction model minimizes **error estimated by DR estimator**

$$\mathcal{L}_{\mathrm{r}} = \sum_{u,i \in \mathcal{D}} \left(\frac{o_{u,i}}{\hat{p}_{u,i}} e_{u,i} + (1 - \frac{o_{u,i}}{\hat{p}_{u,i}}) \hat{e}_{u,i} \right)$$



Error imputation model **minimizes the squared deviation**

$$\mathcal{L}_{\mathrm{e}} = \sum_{u,i \in \mathcal{O}} \frac{(\hat{e}_{u,i} - e_{u,i})^2}{\hat{p}_{u,i}}$$

Analysis of DR Estimator

Bias	$$\begin{array}{c\|c\|c} \mathcal{E}_{\mathrm{EIB}} & \mathcal{E}_{\mathrm{IPS}} & \mathcal{E}_{\mathrm{DR}} \\ \hline \left\| \sum_{u,i \in \mathcal{D}} \frac{(1-p_{u,i})\delta_{u,i}}{\|\mathcal{D}\|} \right\| & \left\| \sum_{u,i \in \mathcal{D}} \frac{\Delta_{u,i} e_{u,i}}{\|\mathcal{D}\|} \right\| & \left\| \sum_{u,i \in \mathcal{D}} \frac{\Delta_{u,i}\delta_{u,i}}{\|\mathcal{D}\|} \right\| \end{array}$$
Tail bound	$$\left\|\mathcal{E}_{\mathrm{DR}} - \mathbb{E}_{\mathbf{O}}[\mathcal{E}_{\mathrm{DR}}]\right\| \leq \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2\|\mathcal{D}\|^2} \sum_{u,i \in \mathcal{D}} \left(\frac{\delta_{u,i}}{\hat{p}_{u,i}}\right)^2}$$
Generalization bound	$$\mathcal{E}_{\mathrm{DR}}(\hat{\mathbf{R}}^{\ddagger}, \mathbf{R}^{o}) + \underbrace{\sum_{u,i \in \mathcal{D}} \frac{\|\Delta_{u,i}\delta^{\ddagger}_{u,i}\|}{\|\mathcal{D}\|}}_{\text{Bias Term}} + \underbrace{\sqrt{\frac{\log\left(\frac{2\|\mathcal{H}\|}{\eta}\right)}{2\|\mathcal{D}\|^2} \sum_{u,i \in \mathcal{D}} \left(\frac{\delta^{\S}_{u,i}}{\hat{p}_{u,i}}\right)^2}}_{\text{Variance Term}}$$

Bias of DR Estimator

Lemma (Bias of DR Estimator)

Given imputed errors $\hat{\mathbf{E}}$ *and learned propensities* $\hat{\mathbf{P}}$ *with* $\hat{p}_{u,i} > 0$ *for all user-item pairs, the bias of the DR estimator is*

$$\text{Bias}(\mathcal{E}_{\mathrm{DR}}) = \frac{1}{|\mathcal{D}|}\left|\sum_{u,i\in\mathcal{D}} \Delta_{u,i}\delta_{u,i}\right|$$

where $\Delta_{u,i} = \frac{\hat{p}_{u,i} - p_{u,i}}{\hat{p}_{u,i}}$ *and* $\delta_{u,i} = e_{u,i} - \hat{e}_{u,i}$.

Corollary (Double Robustness)

The DR estimator is unbiased when either imputed errors $\hat{\mathbf{E}}$ *or learned propensities* $\hat{\mathbf{P}}$ *are accurate for all user-item pairs.*

Tail Bound of DR Estimator

Lemma (Tail Bound of DR Estimator)

Given imputed errors $\hat{\mathbf{E}}$ *and learned propensities* $\hat{\mathbf{P}}$*, for any prediction matrix* $\hat{\mathbf{R}}$*, with probability* $1-\eta$*, the deviation of the DR estimator from its expectation has the following tail bound*

$$\left|\mathcal{E}_{\mathrm{DR}} - \mathbb{E}_{\mathbf{O}}[\mathcal{E}_{\mathrm{DR}}]\right| \leq \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2|\mathcal{D}|^2}\sum_{u,i\in\mathcal{D}}\left(\frac{\delta_{u,i}}{\hat{p}_{u,i}}\right)^2}.$$

Corollary (Tail Bound Comparison)

Suppose imputed errors $\hat{\mathbf{E}}$ *are such that* $0 \leq \hat{e}_{u,i} \leq 2e_{u,i}$ *for* $u, i \in \mathcal{D}$*, then for any learned propensities* $\hat{\mathbf{P}}$*, the tail bound of the DR estimator will be lower than that of the IPS estimator.*

Generalization Bound

Theorem (Generalization Bound)

For any finite hypothesis space $\mathcal{H}$ of prediction matrices, with probability $1-\eta$, the prediction inaccuracy $\mathcal{P}(\hat{\mathbf{R}}^{\ddagger}, \mathbf{R}^{f})$ of the optimal prediction matrix using the DR estimator with imputed errors $\hat{\mathbf{E}}$ and learned propensities $\hat{\mathbf{P}}$ has the upper bound

$$\mathcal{E}_{\mathrm{DR}}(\hat{\mathbf{R}}^{\ddagger}, \mathbf{R}^{o}) + \underbrace{\sum_{u,i\in\mathcal{D}} \frac{|\Delta_{u,i}\delta^{\ddagger}_{u,i}|}{|\mathcal{D}|}}_{\text{Bias Term}} + \underbrace{\sqrt{\frac{\log\left(\frac{2|\mathcal{H}|}{\eta}\right)}{2|\mathcal{D}|^2} \sum_{u,i\in\mathcal{D}} \left(\frac{\delta^{\S}_{u,i}}{\hat{p}_{u,i}}\right)^2}}_{\text{Variance Term}},$$

where $\delta^{\S}_{u,i} = e^{\S}_{u,i} - \hat{e}^{\S}_{u,i}$ is the error deviation corresponding to the prediction matrix $\hat{\mathbf{R}}^{\S} = \operatorname{argmax}_{\hat{\mathbf{R}}^{h}\in\mathcal{H}} \left\{ \sum_{u,i\in\mathcal{D}} \left(\frac{\delta^{h}_{u,i}}{\hat{p}_{u,i}}\right)^2 \right\}$.

Experiments

- MAE and MSE when test on MAR ratings

	Coat		Yahoo	
	MAE	MSE	MAE	MSE
MF	0.920	1.257	1.154	1.891
PMF	0.903	1.239	1.103	1.709
CPT-v	0.969	1.441	0.770	1.115
MF-HI	0.922	1.261	1.158	1.905
MF-MNAR	0.884	1.214	1.177	2.175
MF-IPS	0.860	1.093	0.810	0.989
MF-JL	0.866	1.136	0.899	1.256
MF-DR-JL	**0.778**	**0.990**	**0.747**	**0.966**

* MF-JL and MF-DR-JL are the proposed approaches.

Experiments

- Estimation bias and standard deviation using synthetic data under MSE

	EIB	IPS	SNIPS	NCIS	DR
ONE	22.8±1.8	20.7±1.8	20.7±1.8	26.0±1.7	**9.9±0.9**
FOUR	64.5±1.7	66.8±1.8	66.8±1.8	84.0±1.8	**24.1±0.6**
ROT	18.4±0.3	18.5±0.3	18.5±0.2	23.1±0.2	**10.3**±0.2
SKEW	15.7±0.5	14.8±0.7	14.9±0.5	17.8±0.4	**10.1±0.3**
CRS	18.6±0.3	16.1±0.5	16.2±0.3	20.7±0.2	**9.0±0.1**

Take Away

- Missing ratings are **not always missing at random**
- **Accurate estimation** of the prediction error on MNAR ratings improves **generalization and performance**
- Doubly robust estimator **often gives more accurate** estimation
- **Joint learning** of rating prediction and error imputation achieves further **improvements**

Poster: Today @ Pacific Ballroom **#217**

Thanks for your time!
Questions?

Appendix

Missing At Random and Missing Not At Random

Missing ratings are *missing at random* (MAR), i.e., the probability of observing the indicator matrix only depends on the observed ratings [1]

$$p(\mathbf{O}|\mathbf{R}, \mathbf{X}) = p(\mathbf{O}|\mathbf{R}^o)$$

Missing ratings are *missing not at random* (MNAR), e.g., the probability of a rating being missing depends on its value [2]

$$p(\mathbf{O}|\mathbf{R}, \mathbf{X}) \neq p(\mathbf{O}|\mathbf{R}^o)$$

Appendix

Table: Inaccuracy of rating prediction on MAR test ratings.

	Coat		Yahoo	
	MAE	MSE	MAE	MSE
FM	0.911	1.252	1.154	1.891
NFM	0.888	1.218	1.001	1.488
FM-IPS	0.853	1.086	0.810	0.989
NFM-IPS	0.832	1.065	0.798	0.979
FM-JL	0.859	1.129	1.032	1.528
NFM-JL	0.838	1.114	1.016	1.509
FM-DR-JL	0.775	0.986	0.747	0.966
NFM-DR-JL	**0.756**	**0.967**	**0.736**	**0.957**

* The bottom four rows show the proposed approaches.

Table: Inaccuracy of rating prediction on MNAR test ratings.

	Amazon		Movie	
	MSE	MSE-SNIPS	MSE	MSE-SNIPS
MF	0.949	0.931	0.803	0.793
PMF	0.969	0.911	0.824	0.773
CPT-v	1.277	1.236	1.235	1.180
MF-HI	0.964	0.935	0.812	0.803
MF-MNAR	0.943	0.913	0.803	0.764
MF-IPS	0.956	0.924	0.819	0.780
MF-JL	**0.868**	0.851	**0.767**	0.756
MF-DR-JL	0.871	**0.844**	0.782	**0.745**

* MF-JL and MF-DR-JL are the proposed approaches.