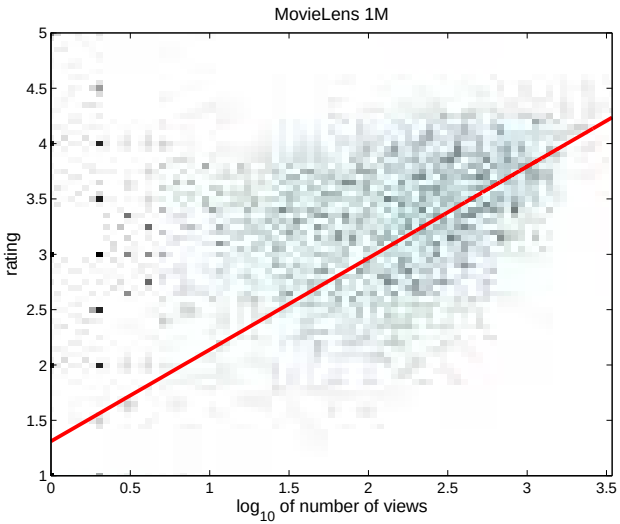
Learning from Missing Data Using Selection Bias in Movie Recommendation

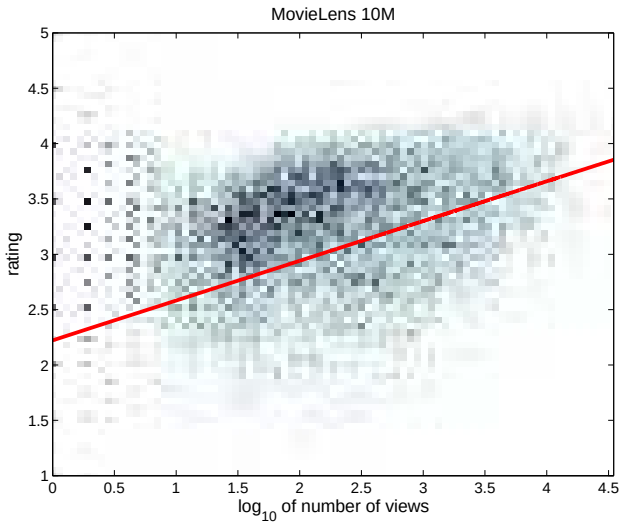
Claire Vernade & Olivier Cappé

ALICIA Meeting – May, 26th and 27th 2015

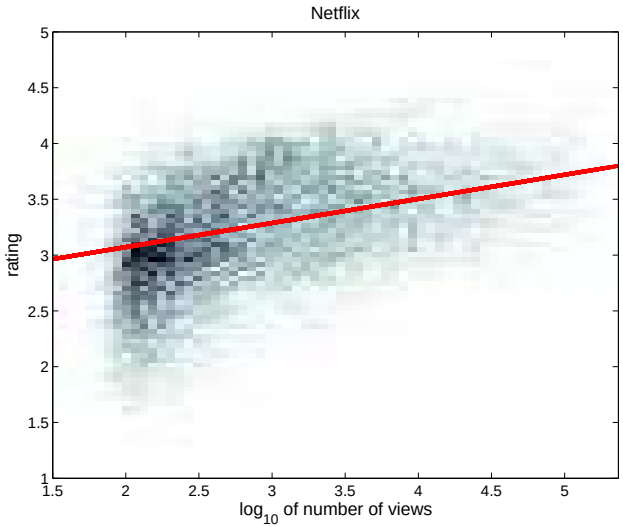
Observing Movie Recommendation data



(a) MovieLens1M



(b) MovieLens10M



(c) Netflix

Investigating the correlation

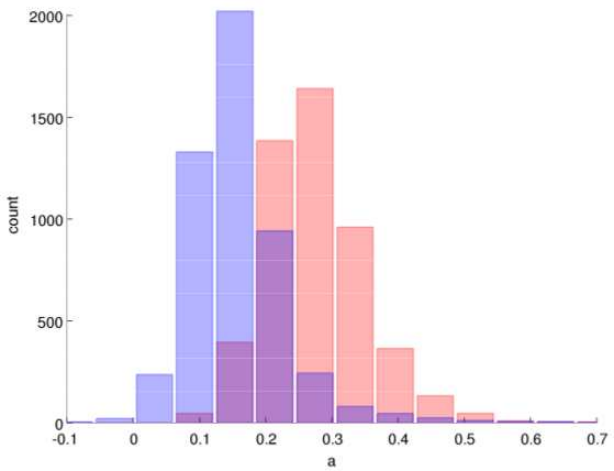


Figure: Histogram of slopes *a* estimated on 5,000 indepedent random subsets of size 100; for MovieLens 10M (red) and Netflix (blue).

Data	ML1M	ML10M	Netflix
full	0.36	0.16	0.09
subsets (100)	-	0.27	0.15

Table: Slope *a* estimated by weighted TLS on the thre complete datasets MovieLens 1M, MovieLens 10M an Netflix (see Fig. 2) and median slopes estimated on random subsets if MovieLens 10M and Netflix (see Fig. 1).

Modelling Selection Bias

The key idea : introducing two linearly dependant parameters :

- $\theta \in \mathbb{R}^K$ representing the average rating of each movie;
- $\beta \in \mathbb{R}^K$ representing the selection probability so that

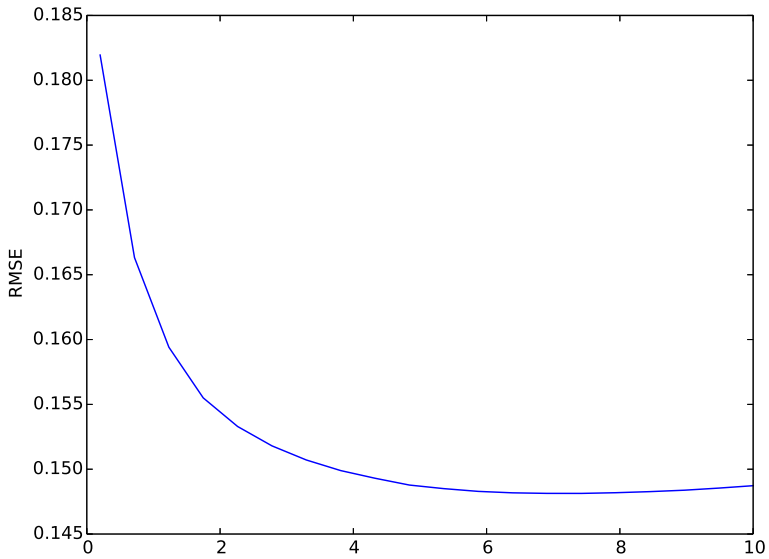
$$p\,(\text{k is selected}) = \frac{e^{\beta}_{k}}{\sum_j e^{\beta_j}}$$

$$\begin{aligned}
\psi^r\left((X_t, Y_t)_{t=1}^n; \theta, \beta\right) &= \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbf{1}\{X_t = k\}\frac{(Y_t - \theta_k)^2}{2\sigma^2} && (1)\\
&\quad - \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbf{1}\{X_t = j\}\beta_k + n\log\left(\sum_{j=1}^{K} e^{\beta_j}\right) && (2)\\
&\quad + r\|\theta - a\beta\|_2^2 && (3)
\end{aligned}$$

Simulated data

Testing the algorithm on simulated data :

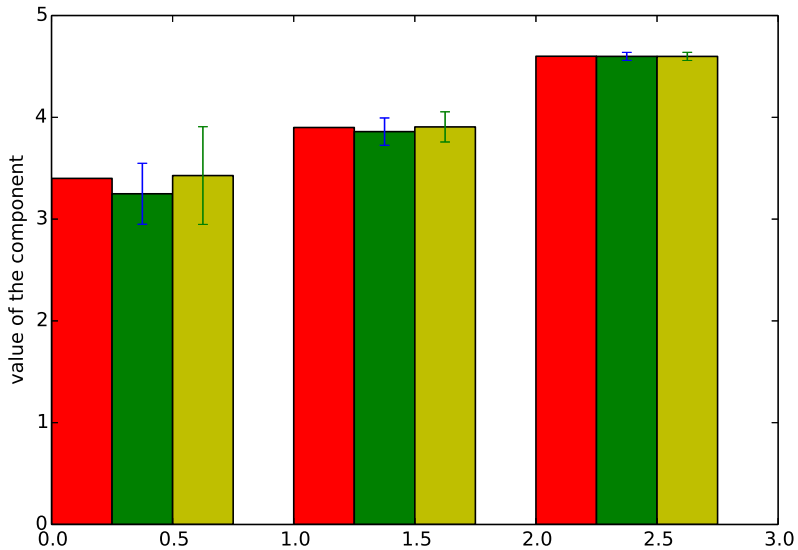
- Simulate data according to the model : noisy β linearly linked to a randomly selected θ
- Choose the regularisation parameter r;
- Observe – and compare with Least Squares – the recovery of an arbitrary parameter θ;
- Compare the efficiency of our estimator with LS as the number of observations grows;
- Observe the influence of an uncertainty on the slope a.



Simulated data

Testing the algorithm on simulated data :

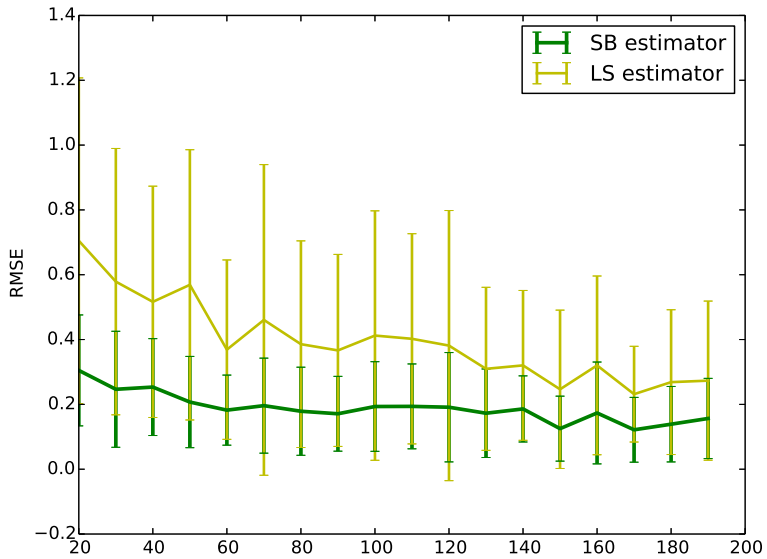
- Simulate data according to the model : noisy β linearly linked to a randomly selected θ
- Choose the regularisation parameter r;
- Observe – and compare with Least Squares – the recovery of an arbitrary parameter θ;
- Compare the efficiency of our estimator with LS as the number of observations grows;
- Observe the influence of an uncertainty on the slope a.



Simulated data

Testing the algorithm on simulated data :

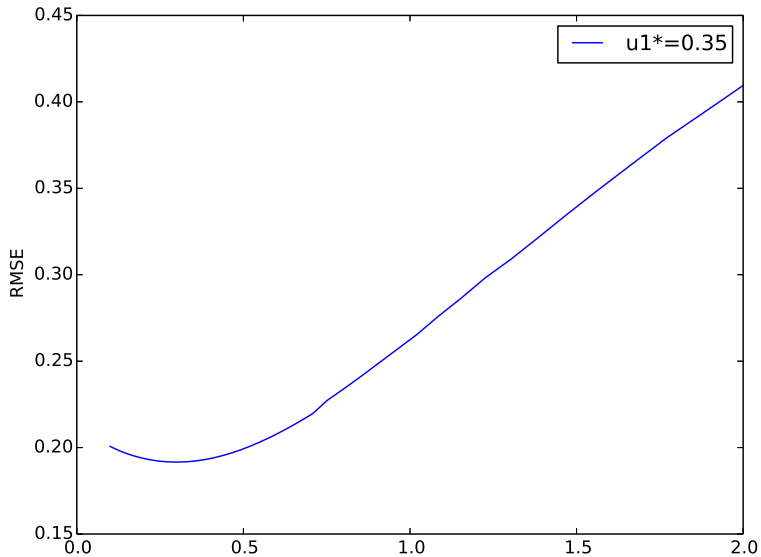
- Simulate data according to the model : noisy β linearly linked to a randomly selected θ
- Choose the regularisation parameter r;
- Observe – and compare with Least Squares – the recovery of an arbitrary parameter θ;
- Compare the efficiency of our estimator with LS as the number of observations grows;
- Observe the influence of an uncertainty on the slope a.



Simulated data

Testing the algorithm on simulated data :

- Simulate data according to the model : noisy β linearly linked to a randomly selected θ
- Choose the regularisation parameter r;
- Observe – and compare with Least Squares – the recovery of an arbitrary parameter θ;
- Compare the efficiency of our estimator with LS as the number of observations grows;
- Observe the influence of an uncertainty on the slope a.



Real Data Experiments

- Neighbourhood methods : similarity based on a rank-25 SVD and cosine distance between user features.
- For each active user :
  1. Evaluate 100 nearest neighbours;
  2. compute estimates for the unobserved ratings with 4 different methods : Selection Bias, similarity-weighted Selection Bias, Least Squares and Weighted Least Squares;
  3. evaluate RMSE and Precision-at-N
- Two averaged metrics over all users : Averaged RMSE and Precision-at-N;
- Two datasets : MovieLens 1M and MovieLens 10M.

Results

MovieLens1M

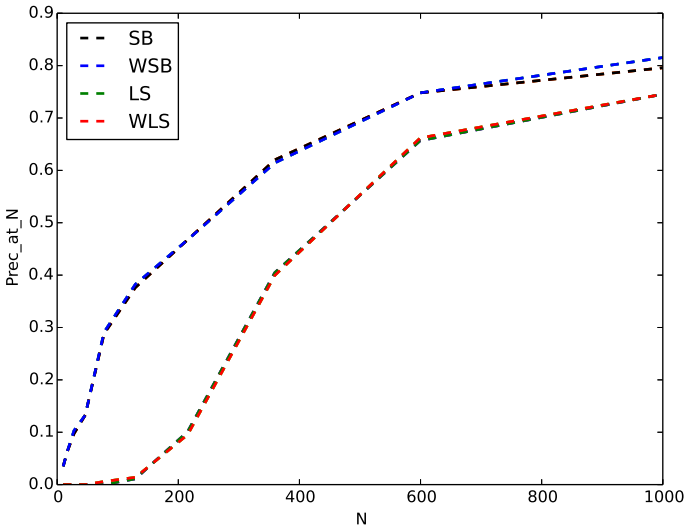


Figure: Precision-at-N as N grows.

	RMSE	P@N ($N = 15$)
$\hat{\theta}(SB)$	**1.272**	**0.063**
$\hat{\theta}(LS)$	1.303	0.0
$\hat{\theta}(WLS)$	1.319	0.0

Table: Results of the experiments on the MovieLens 1 dataset. RMSE and Precision-at-N for $N = 15$.

Conclusions and future work

What we have done

- Investigated the correlation between the average rating of the items and their popularity;
- Proposed a variational model for estimating unobserved ratings using both observations;
- Empirically tested on simulated data the influence of the different quantities introduced (the slope a, the regularisation r, the size of the database N);
- Plugged our estimator and its similarity-weighted version in neighbourhood methods for recommendation and compared it to Least Squares.

What can be done

- Use similar ideas in other frameworks such that Matrix Completion;
- Investigate similarity taking Selection Bias into account;