MMM, LDA & CTM

Yotam cohen - 203551734, Dror Hadas - 302678263

1 Intro

We present a comparison between three models from the field of Topic Modeling:

- MMM - As presented in class, in mode de-novo.
- LDA - Latent Dirichlet Allocation (Paper: David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003), 993-1022. **link**)
- CTM - Correlated Topic Modeling (Paper: David M. Blei and John D. Lafferty. 2005. Correlated topic models. In Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05), Y. Weiss, B. Schölkopf, and J. C. Platt (Eds.). MIT Press, Cambridge, MA, USA, 147-154. link)

2 Results

2.1 Choosing the optimal amount of topics

We used the STM R package (which we use to run the CTM model) in order to choose the most suitable number of exposures. We used the *searchK* model which receives a numeric vector of the topic numbers to compare, and the observational mutations matrix. For each one it runs the CTM model and calculates a number of metrics in order to compare the models:

- *Log-likelihood* - Computed with 'held-out' data (0.2 proportion).
- *Semantic coherency* - This metric is used to assess the quality of the topics. Roughly, it checks that the most probable words of each topic occur together.
- *Exclusivity* - Together with semantic coherency, we want to make sure that the topics differs from one another. This metric uses the usage rate of each word in relation to its usage rate in other topics.

From this, we concluded that 12 and 14 are suitable topic numbers for the model.
We notice that the package *Exclusivity* metric is marked for internal use, as it doesn't have error checking. For that reason, we did not depend on its result and we don't show the result of that computation. It is recommended to use it, or any other implementation, to ensure the correct topics number.

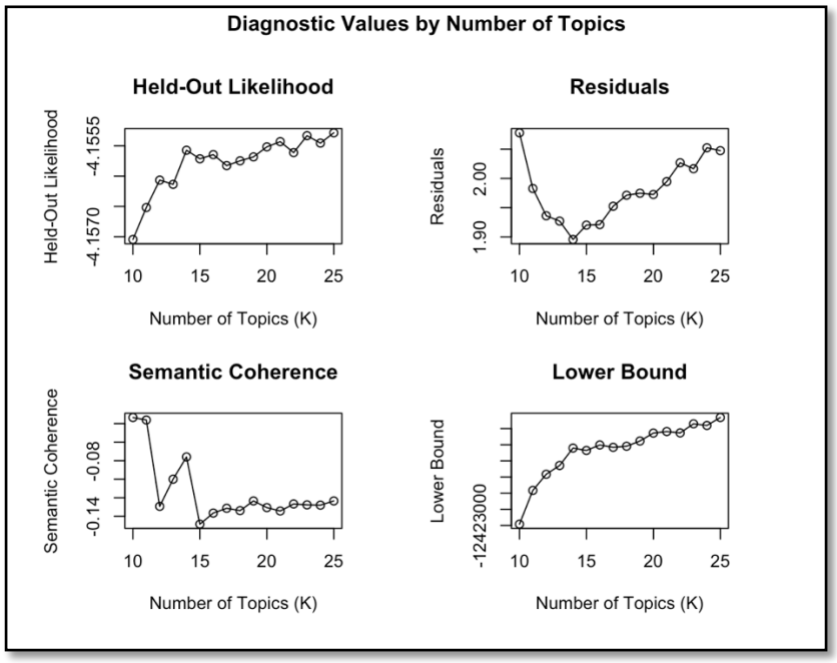

Figure 1. Line plot for each one of the topics number a) Log-likelihood b) Residuals c) Semantic coherence d) Lower Bound.

2.2 Models comparison with 12 topics and 14 topics

Our first and main goal was to perform a comparison between the models based on the provided data, as we couldn't fixate the mutation matrix in either CTM/LDA algorithms we compered the models using 3 different strategies:

2.2.1 Direct log-likelihood comparisons

Method:

1. *Trained each model on all the data.*
2. *Extracted exposure and mutations matrices.*
3. *Calculated LL using our implementation.*

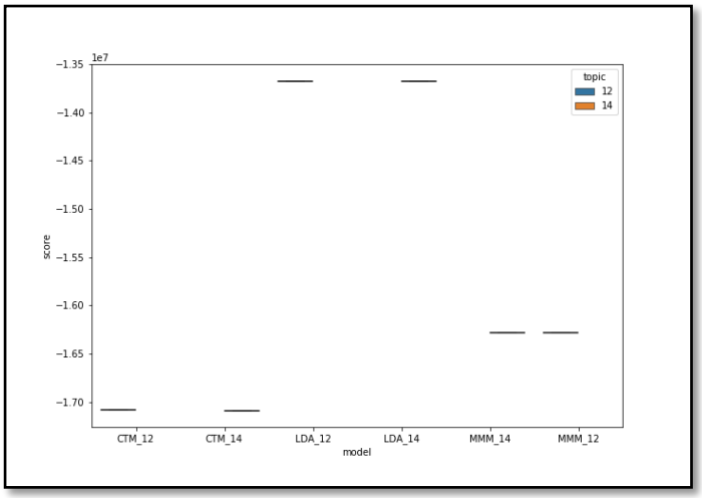

Figure 2. LL of direct testing scores of 6 models LDA, CTM and MMM with 12 or 14 topics.

In order to avoid over-fitting issues, we performed an analysis while training on 22 chromosomes and performing the LL on the held-out chromosome, and managed to receive similar results.

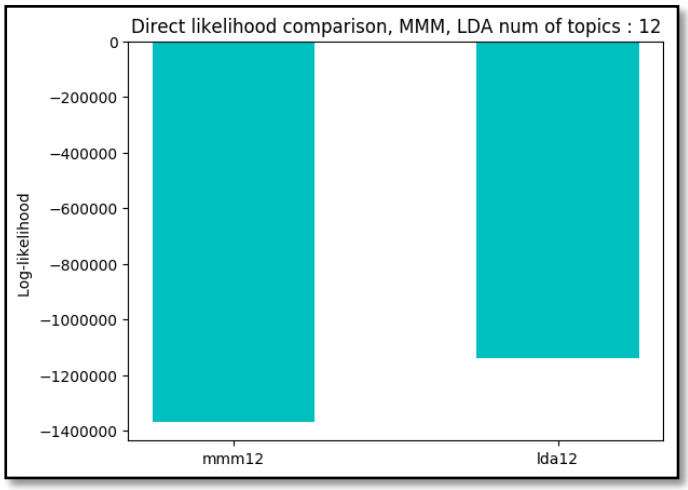

*Figure 2.b LL of direct testing scores of 2 models LDA, MMM with 12, and cross validation of 1 chromosome out. Exact values are: **MMM**: -1366957.59, **LDA** : -1139873.86*

It is possible to see that LDA preforms better in that criteria.

2.2.2 Generalization of samples

Method:

1. *10-fold cross-validation on the samples. train = 504, test = 56.*
2. *Trained each model.*
3. *Refitted with our MMM implementation using the test data.*
4. *Calculated LL using our implementation*

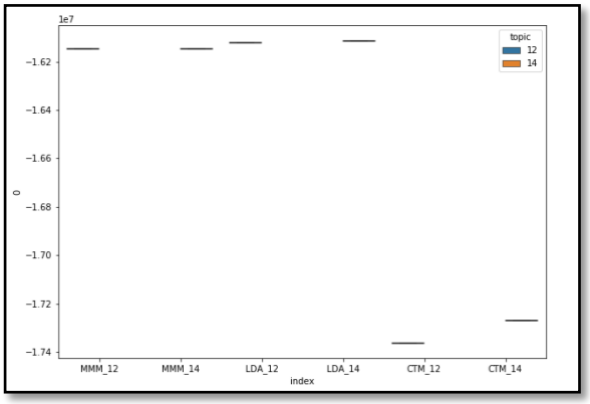

Figure 3. LL scores of 6 models LDA, CTM and MMM with 12 or 14 topics while generalizing new samples.

2.2.3 Cross-validation by chromosomes

Method:

1. *Trained each model on all the data.*
2. *Extracted the mutations matrix.*
3. *Refitted * using our MMM implementation with 23-fold CV for each chromosome.*
4. *Calculated LL using our implementation on the held-out chromosome.*

**mmm_refit was done on the mutations matrix provided to us, and wasn't trained, but only refitted in cross-validation.*

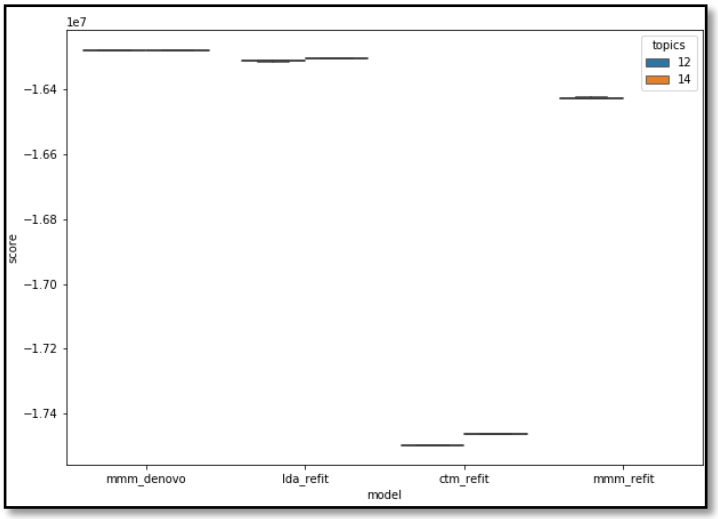

Figure 4. LL scores of 7 models LDA, CTM and MMM with 12 or 14 topics and of MMM after refit.

3 Additional results

As we progressed with the development of the current model, we have collected some additional results and conclusions regarding the data.

3.1 Correlations between the different topics

The CTM model uses a Sigma prior which we can infer the correlations between the topics which were used by the model.

We can see that topics 1 and 5 are highly corelated with each other (Figure 3), but are not related at all to the rest of the topics, as we can see in the neighbors' graph (Figure 4).

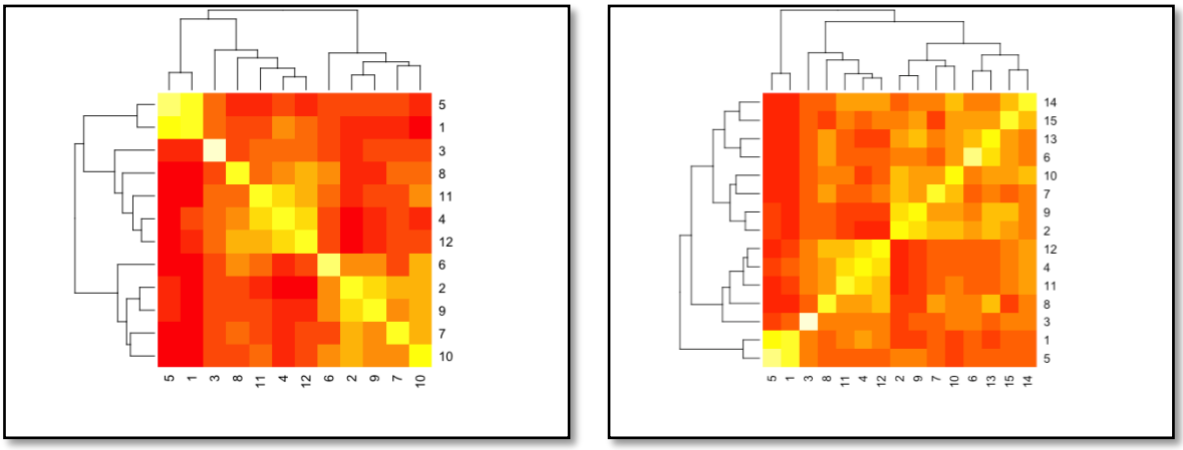

Figure 5. Heatmap correlations between the topics as inferred by the CTM model. Red color stands for a negative correlation between topics and yellow stands for positive ones.

a) correlation of a 12-topic model b) correlation of a 15-topic model.

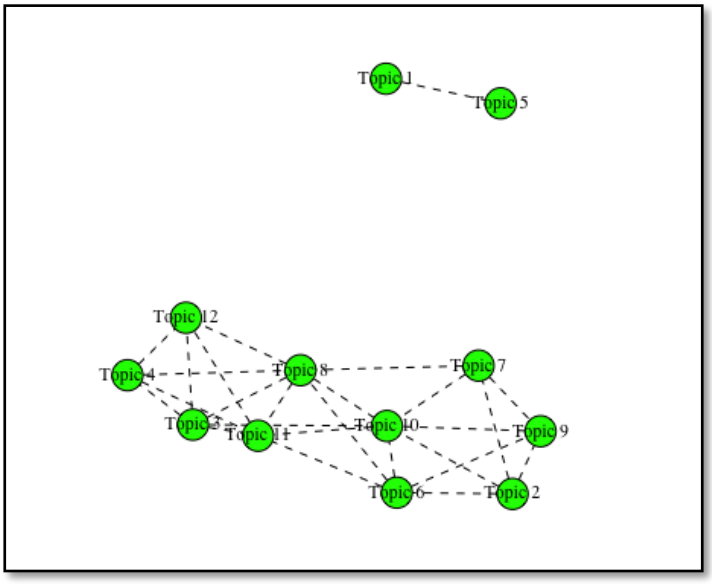

Figure 6. Neighbors' graph between topics.

3.2 PCA and t-SNE of exposure vectors

Next, we sought to check if we can see a separation between different people, using the exposure vectors. For this purpose, we applied 2 methods of dimensionality reduction (PCA & t-SNE). A similar output is yielded by preforming the same on the LDA's output but is not provided here.

In both algorithms, we noticed a strong separation among one of the axes. We wanted to see if the separation is of the same samples, so we applied a person correlation test which shows a strong negative correlation (-0.9249473) between the models X axis. (Figure 7).

As we lacked the metadata to perform a thorough research regarding the meaning of the different exposure and subjects, we have decided to note this as an interesting observation.

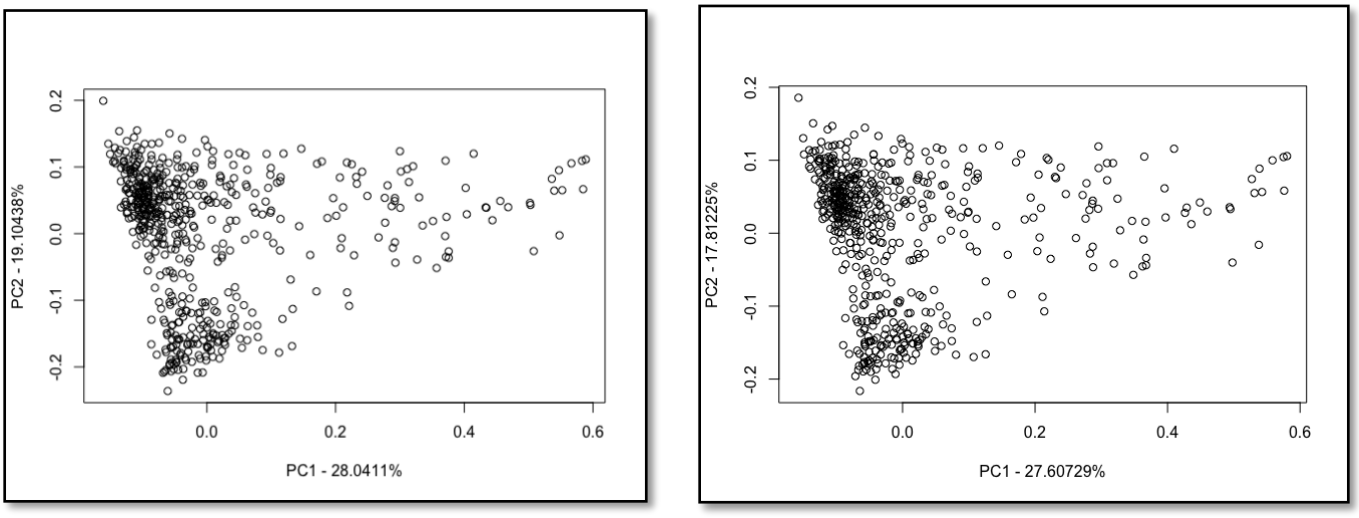

Figure 7. PCA of CTM exposures, 12 topics (left) and 14 topics (right).

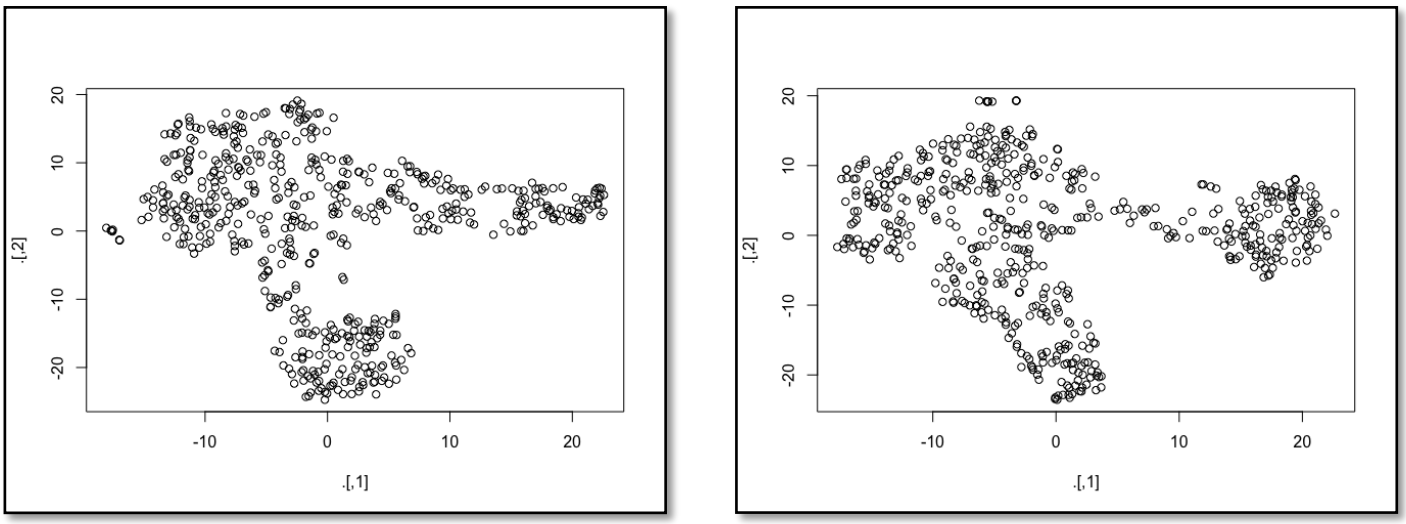

Figure 8. TSNE output of the CTM exposure vectors: 12 topics (left), 14 topics (right)

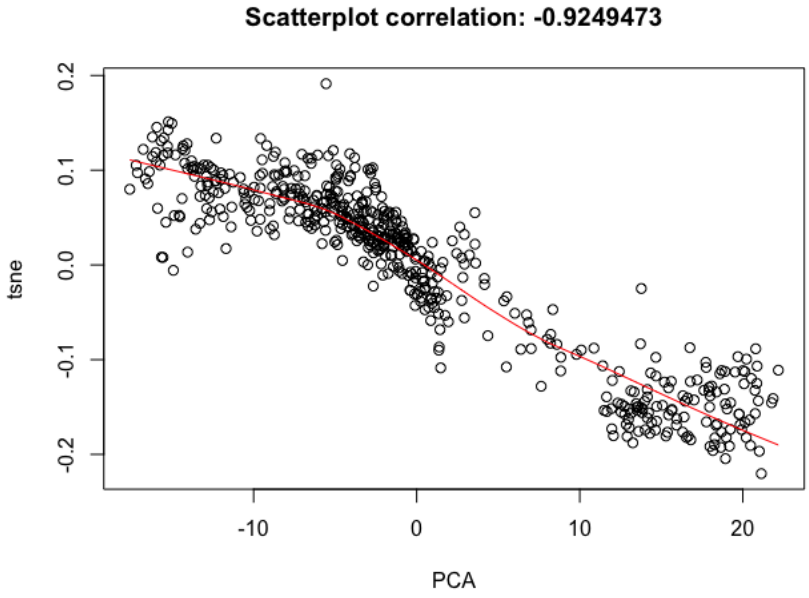

Figure 9. TSNE and PCA axis y correlation, performed on the 12 topics exposure matrix.

4 Methods and resources:

- LDA calculations were made with the package 'topicmodels' for R. Source
- CTM calculations were made with the package 'STM' source. Source
- MMM code is based on Paper provided in class and is provided in Git.
- MMM model **implementation** is based on the class lectures of course 0368.4212 – Biological networks analysis, Tel Aviv university by prof. Roded Sharan & Mr. Itay Sason.

All of the data and code to calculate the models, check the topics number and generate the plots is stored in the provided git repository.