The New York Times

THE Nation.

NATIONAL REVIEW

BREITBART

THE WALL STREET JOURNAL.

Slate

THE ROOT

words connected to news coverage of Black Lives Matter, terrorism, election etc for different newspapers, particularly, ones reflecting bias in coverage

The New York Times

THE Nation.

NATIONAL REVIEW

BREITBART

THE WALL STREET JOURNAL.

Slate

THE ROOT

Black Lives Matter Keeps Getting More Radical — Will the Media Care?

These Americans of good will have been had, with the media's help. Black Lives Matter is an instrument not of justice and reconciliation but rather of violence and revolution. It's shot through with its own form of racism — anti-Semitism — and with authoritarian demands that would not only strip Americans of their constitutional rights but bankrupt our nation and render it vulnerable to its enemies abroad. Doubt me? Read the organization's own words.

GUNS AND GUN CONTROL BLACK LIVES MATTER EDITORIAL AUGUST 1-8, 2016 ISSUE

Black Lives Still Matter

In order to truly ensure that, we will have to confront the broader culture of violence that has long gripped this nation.

RACISM AND DISCRIMINATION BLACK LIVES MATTER EDITORIAL AUGUST 29-SEPTEMBER 5, 2016 ISSUE

What Does Black Lives Matter Want? Now Its Demands Are Clearer Than Ever

After a year of planning, members of the movement have released a comprehensive platform.

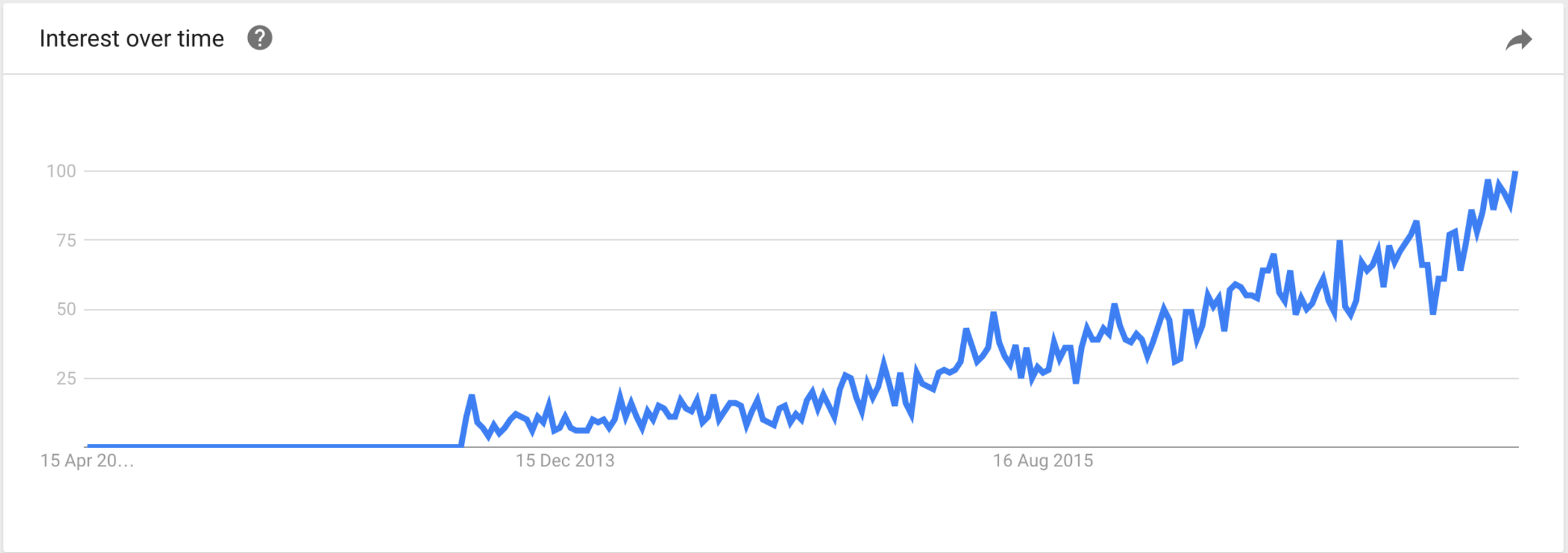
At both the Democratic and Republican conventions last month, there were plenty of indications that conversations strengthened and sustained by the current movement to end antiblack racism have made it to the national stage. The "Mothers of the Movement"—women whose children were killed by police or vigilantes or who died while in police custody—shared their stories at the Democratic National Convention, making the case that their fight for justice would be in good hands with a Clinton presidency. The previous

Black Lives Matter: Radicals Using Moderates to Help Tear America Apart

Black Lives Matter supporters march in New York City, July 10, 2016. (Eduardo Munoz/Reuters)

word2vec

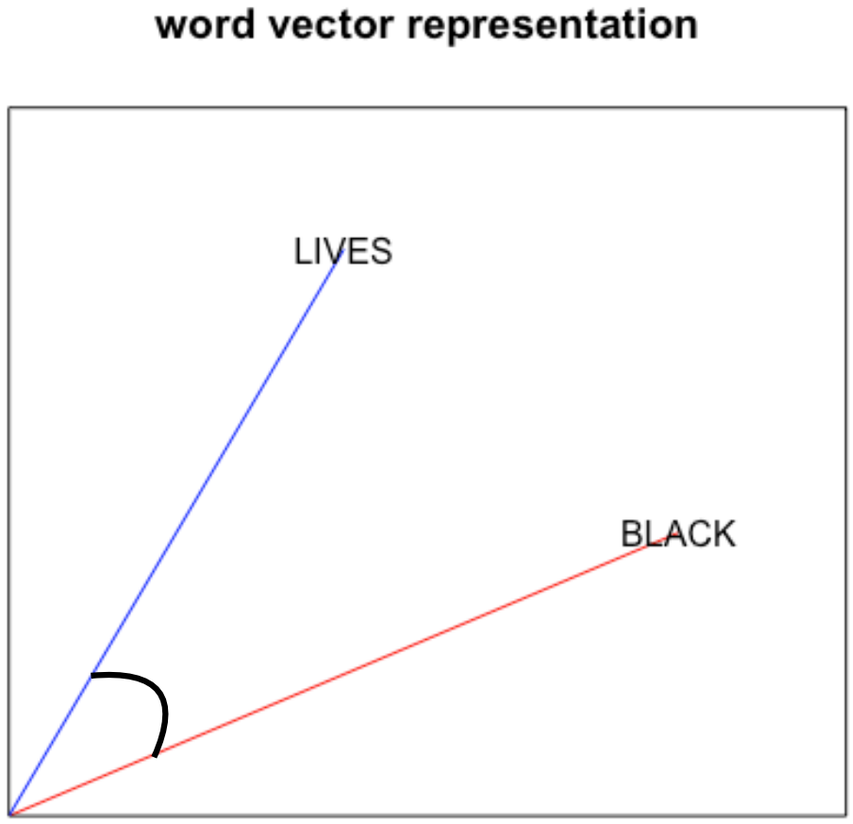
A tool gaining a lot of interest in ML- based Natural Language Processing.



Word Embeddings

presents each word as a vector in some space

What is the cosine of the angle between vec(“black”) and vec(“lives”)



Example of word embedding

Consider the following set of texts

I am speaking
I was eating
I was traveling

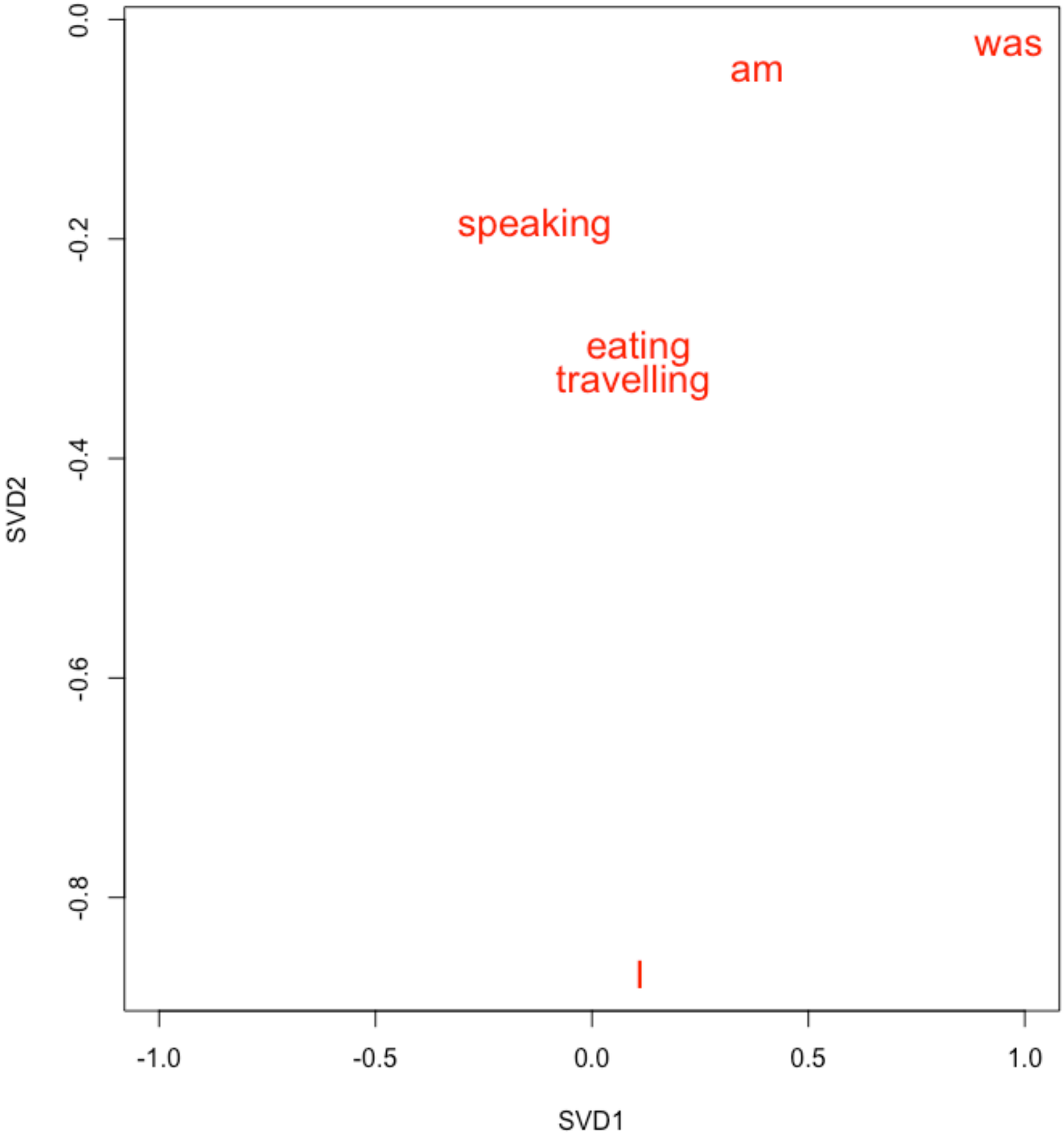
Example of word embedding

Consider the following set of texts

I am speaking
I was eating
I was traveling

Co-occurrence Matrix

	I	am	was	speaking	eating	travelling
I	0	1	2	0	0	0
am	1	0	0	1	0	0
was	2	0	0	0	1	1
speaking	0	1	0	0	0	0
eating	0	0	1	0	0	0
travelling	0	0	1	0	0	0


was
am
speaking
eating
travelling
0.0
-0.2
-0.4
-0.6
-0.8
SVD2
-1.0
-0.5
0.0
0.5
1.0
SVD1

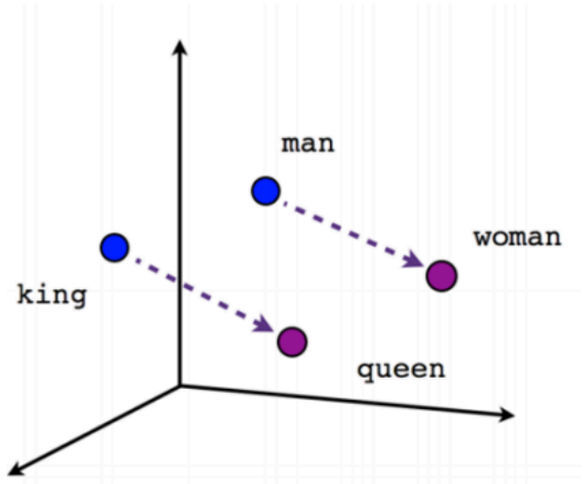
Why word2vec

word2vec usually projects on large
dimensional vector space - 50 or 100 in general

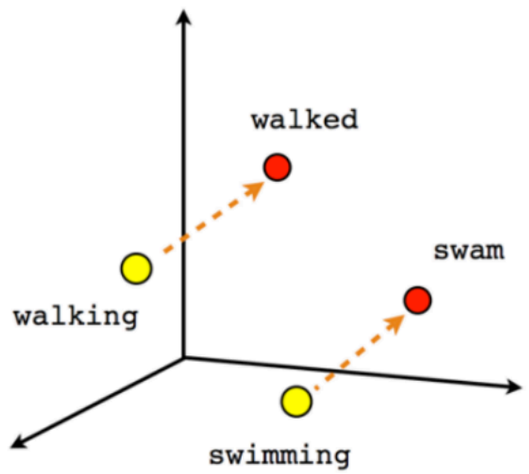
Why word2vec

word2vec usually projects on large
dimensional vector space - 50 or 100 in general

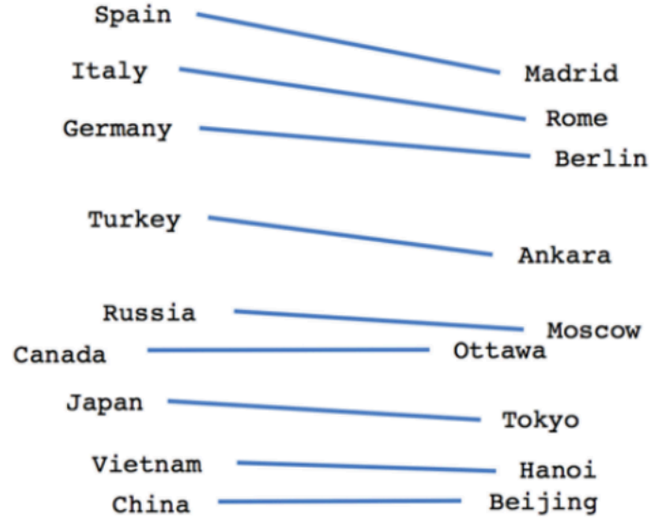
t-SNE projection of these high word2vec vectors



Male-Female



Verb tense



Country-Capital

Mikolov et al 2013

vec(“king”) - vec(“man”) + vec(“woman”) = vec(“queen”)

Consider a word w_t

I	**am**	**giving**	**a**	**talk**
w_{t-2}	w_{t-1}	w_t	w_{t+1}	w_{t+2}

Consider a word w_t

I	**am**	**giving**	**a**	**talk**
w_{t-2}	w_{t-1}	w_t	w_{t+1}	w_{t+2}

We maximize the following objective function

$$J = \frac{1}{T} \sum_{t=1}^{T} \sum_{-n \leq j \leq n, j \neq 0} \log p\left(w_{t+j} | w_t\right)$$

Consider a word w_t

I	**am**	**giving**	**a**	**talk**
w_{t-2}	w_{t-1}	w_t	w_{t+1}	w_{t+2}

We maximize the following objective function

$$J = \frac{1}{T}\sum_{t=1}^{T}\sum_{-n\leq j\leq n, j\neq 0} \log p\left(w_{t+j}|w_t\right)$$

$$p\left(w_{t+j}|w_t\right) = \frac{exp\left(v_{w_t}^T u_{w_{t+j}}\right)}{\sum_{w_i\sim V} exp\left(v_{w_t}^T u_{w_i}\right)}$$

The parameters: $\left(v_w, u_w \forall w\right)$

7020 Articles scraped from the Nation
between Jan 14 and Mar 16.
Data was missing in August 2014 and 2015.

Keywords:

black, lives, matter, police, brutality, racism, crime, violence, laquan, mcdonald, trump, clinton, terrorism

```
> nearest_to(model2, model2[["black"]], 20)
            black            latino             white           unarmed
     2.220446e-16      3.238422e-01      3.296694e-01      3.950784e-01
    subordination            blacks         vigilantes             lives
     4.032799e-01      4.125458e-01      4.227586e-01      4.229431e-01
         teenager            african         supremacy            latinos
     4.289137e-01      4.309424e-01      4.324514e-01      4.381511e-01
            young              men             racial    criminalization
     4.412078e-01      4.485662e-01      4.556035e-01      4.559089e-01
           dylann            jackets           ferguson disproportionately
     4.573308e-01      4.578751e-01      4.581132e-01      4.618305e-01
```

top words - **Clinton**

hillary	clinton's	sanders	candidacy	bernie	presumptive
0.9154660	0.8522816	0.7526696	0.6760095	0.6666568	0.6490119
hillary's	vt	sanders's	rodham	campaign's	spar
0.6068201	0.5969407	0.5957619	0.5955374	0.5931926	0.5875940
walters	thorny	democratic	dfa	candidate	woodruff
0.5790676	0.5735087	0.5717317	0.5544882	0.5514192	0.5449966
nomination	unbeatable				
0.5447019	0.5443936				

… the palestinian authority pa and hamas spar including health public works ….

……*bernie sanders and hillary clinton spar over which one is more progressive……*

…. from the public by having them spar mostly on weekends or holidays ……

top words - **Clinton**

hillary	clinton's	sanders	candidacy	bernie	presumptive
0.9154660	0.8522816	0.7526696	0.6760095	0.6666568	0.6490119
hillary's	vt	sanders's	rodham	campaign's	spar
0.6068201	0.5969407	0.5957619	0.5955374	0.5931926	0.5875940
walters	thorny	democratic	dfa	candidate	woodruff
0.5790676	0.5735087	0.5717317	0.5544882	0.5514192	0.5449966
nomination	unbeatable				
0.5447019	0.5443936				

… the palestinian authority pa and hamas spar including health public works ….

……*bernie sanders and hillary clinton spar over which one is more progressive……*

…. from the public by having them spar mostly on weekends or holidays ……

top words - **matter**

lives	black	question	devoid
0.6518010	0.5320571	0.5103602	0.4554311
domestically	matters	blacklivesmatter	longer
0.4390956	0.4380126	0.4345750	0.4343148
there's	exists	clear	coincidence
0.4306680	0.4301122	0.4292051	0.4144670
hashtags	makes	dismissively	surmised
0.4135828	0.4037018	0.3983225	0.3979131
racist	supremacy	exist	imaginable
0.3926556	0.3920765	0.3852060	0.3838286

three occurrences

…….the genre of drone videos grows increasingly popular domestically weddings sporting events……

…….about the values that guide this country as it engages domestically and internationally……

…….religious rights seem to be a matter of rare consensus both domestically and internationally……

CorShrink

on word2vec cosine similarities

We obtain the **word2vec** cosine similarity values between any two words of interest - say ***black*** and ***lives***. We treat cosine similarity as correlation.

We obtain the **word2vec** cosine similarity values between any two words of interest - say ***black*** and ***lives***. We treat cosine similarity as correlation.

Convert the cosine similarities into **Fisher z-scores**

$$\rho = \frac{1}{2}\log(\frac{1+r}{1-r})$$

We obtain the **word2vec** cosine similarity values between any two words of interest - say ***black*** and ***lives***. We treat cosine similarity as correlation.

Convert the cosine similarities into **Fisher z-scores**

$$\rho = \frac{1}{2}\log(\frac{1+r}{1-r})$$

Obtain **bootstrap** standard error for the Fisher z-scores by sampling texts with replacement, pooling them, and running **word2vec** on those models. We chose bootstrap size **100**.

We obtain the **word2vec** cosine similarity values between any two words of interest - say ***black*** and ***lives***. We treat cosine similarity as correlation.

Convert the cosine similarities into **Fisher z-scores**

$$\rho = \frac{1}{2}\log(\frac{1+r}{1-r})$$

Obtain **bootstrap** standard error for the Fisher z-scores by sampling texts with replacement, pooling them, and running **word2vec** on those models. We chose bootstrap size **100**.

For each word, say ***black***, we record the Fisher z-scores for the top **1000** words and their Bootstrap standard errors, and apply **ash** on them.

We obtain the **word2vec** cosine similarity values between any two words of interest - say ***black*** and ***lives***. We treat cosine similarity as correlation.

Convert the cosine similarities into **Fisher z-scores**

$$\rho = \frac{1}{2}\log(\frac{1+r}{1-r})$$

Obtain **bootstrap** standard error for the Fisher z-scores by sampling texts with replacement, pooling them, and running **word2vec** on those models. We chose bootstrap size **100**.

For each word, say ***black***, we record the Fisher z-scores for the top **1000** words and their Bootstrap standard errors, and apply **ash** on them.

Inverse transform the posterior mean of the Fisher z-transform from ash output to get a vector of **shrunk correlations**

We generate **rankings** based on the **shrunk cosine similarities**.

top words - **Clinton** (before ***CorShrink*** adjustment)

hillary	clinton's	sanders	candidacy	bernie	presumptive
0.9154660	0.8522816	0.7526696	0.6760095	0.6666568	0.6490119
hillary's	vt	sanders's	rodham	campaign's	spar
0.6068201	0.5969407	0.5957619	0.5955374	0.5931926	0.5875940
walters	thorny	democratic	dfa	candidate	woodruff
0.5790676	0.5735087	0.5717317	0.5544882	0.5514192	0.5449966
nomination	unbeatable				
0.5447019	0.5443936				

top words - **Clinton** (after ***CorShrink*** adjustment)

hillary	clinton's	sanders	bernie	candidacy	presumptive
0.9054996	0.8522804	0.7381915	0.6333450	0.6296040	0.5796187
campaign's	sanders's	hillary's	democratic	vt	candidate
0.5780555	0.5727214	0.5657258	0.5581314	0.5469533	0.5305830
walters	campaign	nomination	woodruff	vermont	rodham
0.5274639	0.5128978	0.5107737	0.5080026	0.5028814	0.4974063
insurgent	she				
0.4938642	0.4894352				

spar ranks **66th** after *CorShrink* is applied

top words - **matter** (before ***CorShrink*** adjustment)

lives	black	question	devoid
0.6518010	0.5320571	0.5103602	0.4554311
domestically	matters	blacklivesmatter	longer
0.4390956	0.4380126	0.4345750	0.4343148
there's	exists	clear	coincidence
0.4306680	0.4301122	0.4292051	0.4144670
hashtags	makes	dismissively	surmised
0.4135828	0.4037018	0.3983225	0.3979131
racist	supremacy	exist	imaginable
0.3926556	0.3920765	0.3852060	0.3838286

top words - **matter** (after ***CorShrink*** adjustment)

lives	question	black	clear
0.6046905	0.4454659	0.4233648	0.3447135
longer	exists	blacklivesmatter	there's
0.3429932	0.3402007	0.3374417	0.3371693
makes	matters	racist	movement
0.3321932	0.3290786	0.3265337	0.3264511
coincidence	no	devoid	imaginable
0.3255307	0.3239187	0.3230350	0.3229040
supremacy	mobilization	surmised	exist
0.3223244	0.3205559	0.3200350	0.3195980

domestically ranks **180th** after *CorShrink* is applied

CorShrink

Application to GTEx tissue similarities

Problem Specification

Problem Specification

GTEx (Genotype Tissue Expression) project has collected reads expression across genes for samples coming from 51 different tissues and 2 cell lines.

Problem Specification

GTEx (Genotype Tissue Expression) project has collected reads expression across genes for samples coming from 51 different tissues and 2 cell lines.

However not all persons have contributed all tissues. Some tissues have high contributions from people while some others have low contributions.

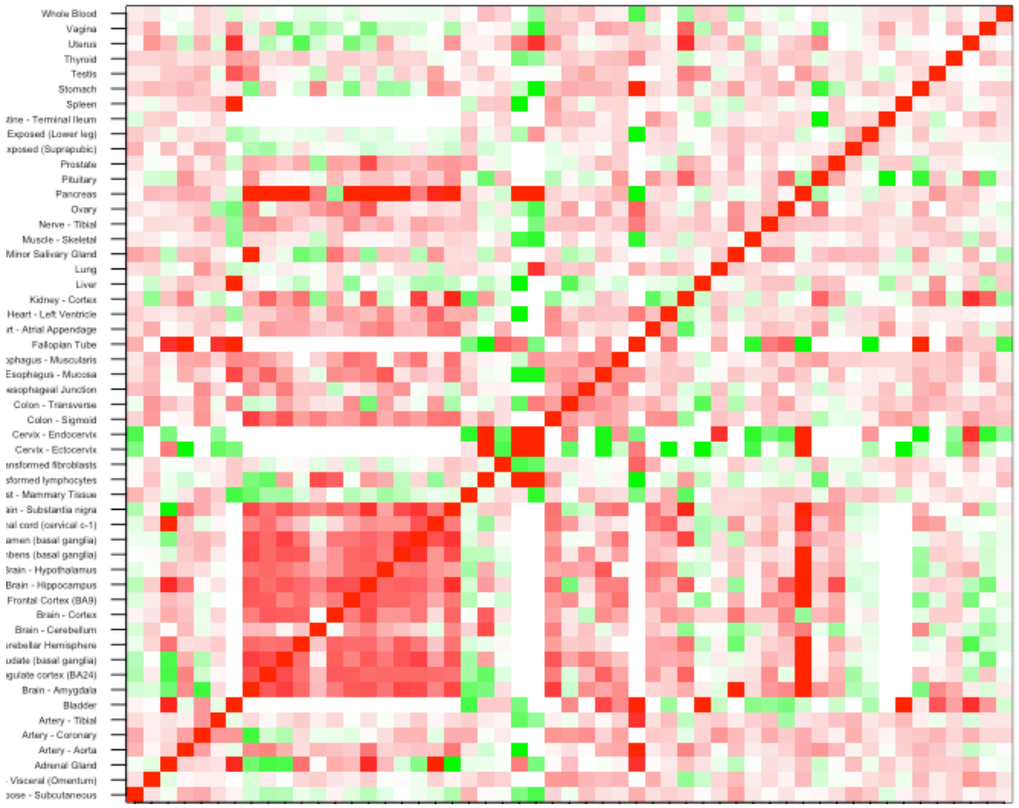
Problem Specification

GTEx (Genotype Tissue Expression) project has collected reads expression across genes for samples coming from 51 different tissues and 2 cell lines.

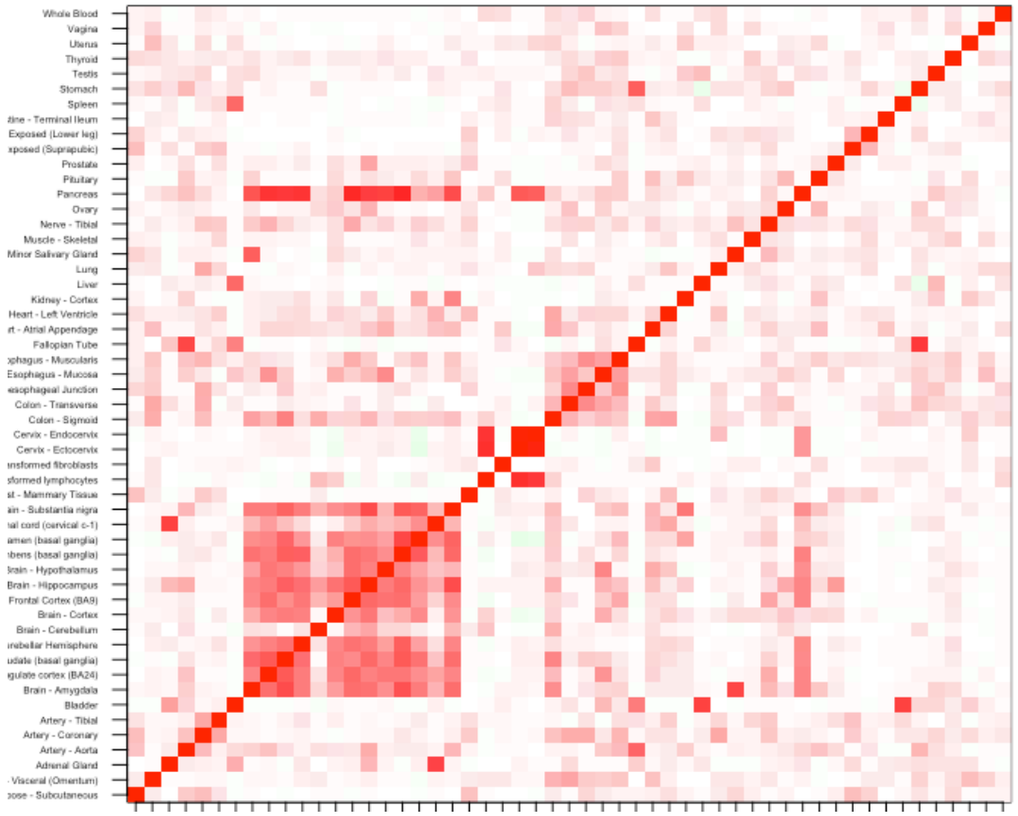
However not all persons have contributed all tissues. Some tissues have high contributions from people while some others have low contributions.

When correlation between gene expression for two tissues for a single gene is measured, it may be affected due to small number of common subjects who have contributed both the tissues.

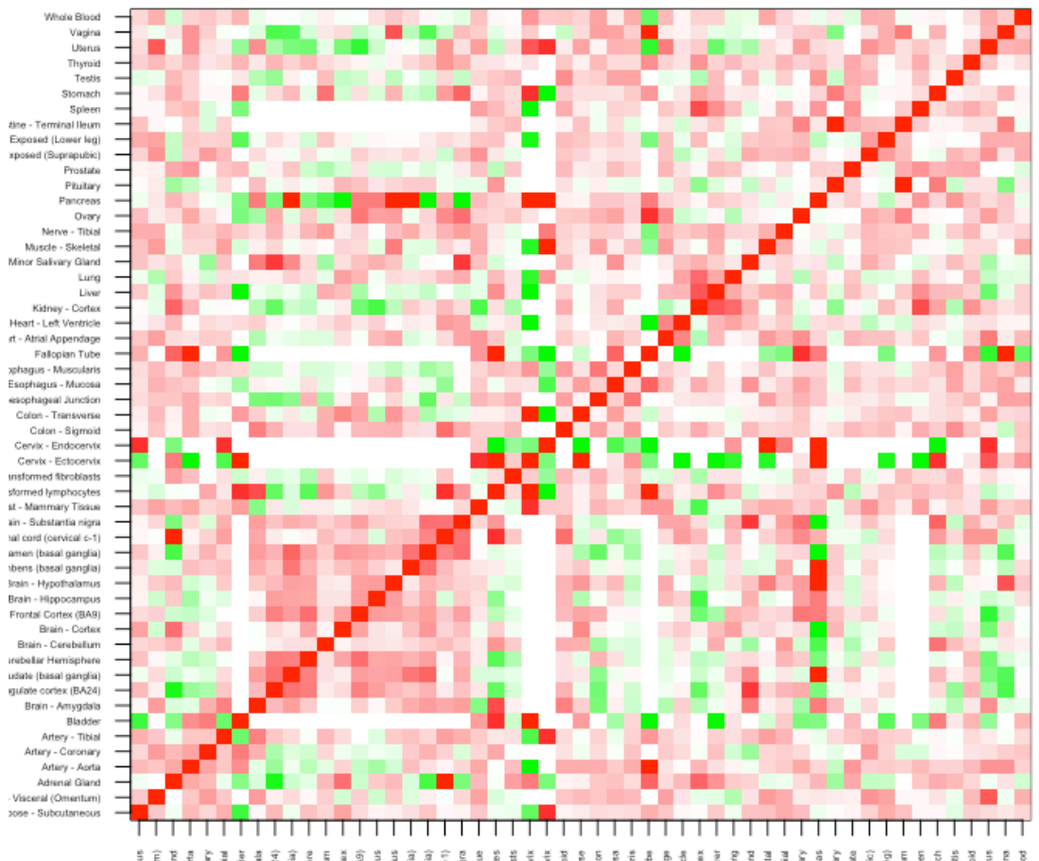
Corr mat: ENSG00000000971



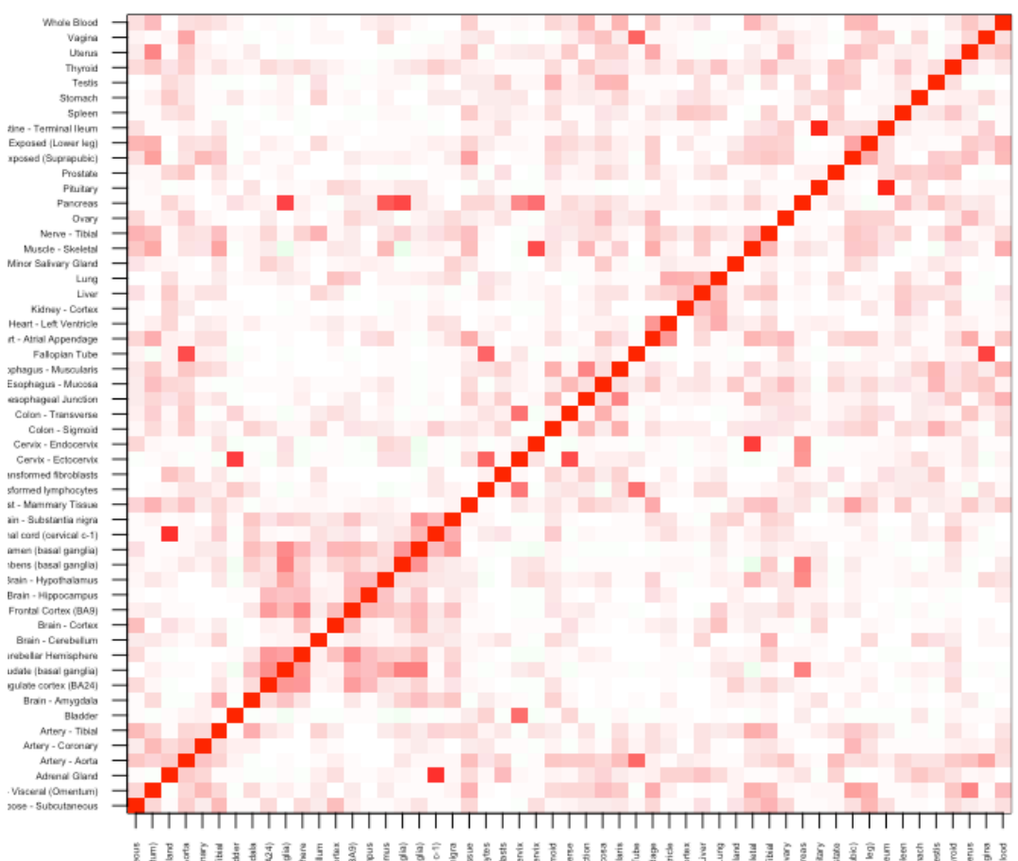
CorShrink mat: ENSG00000000971



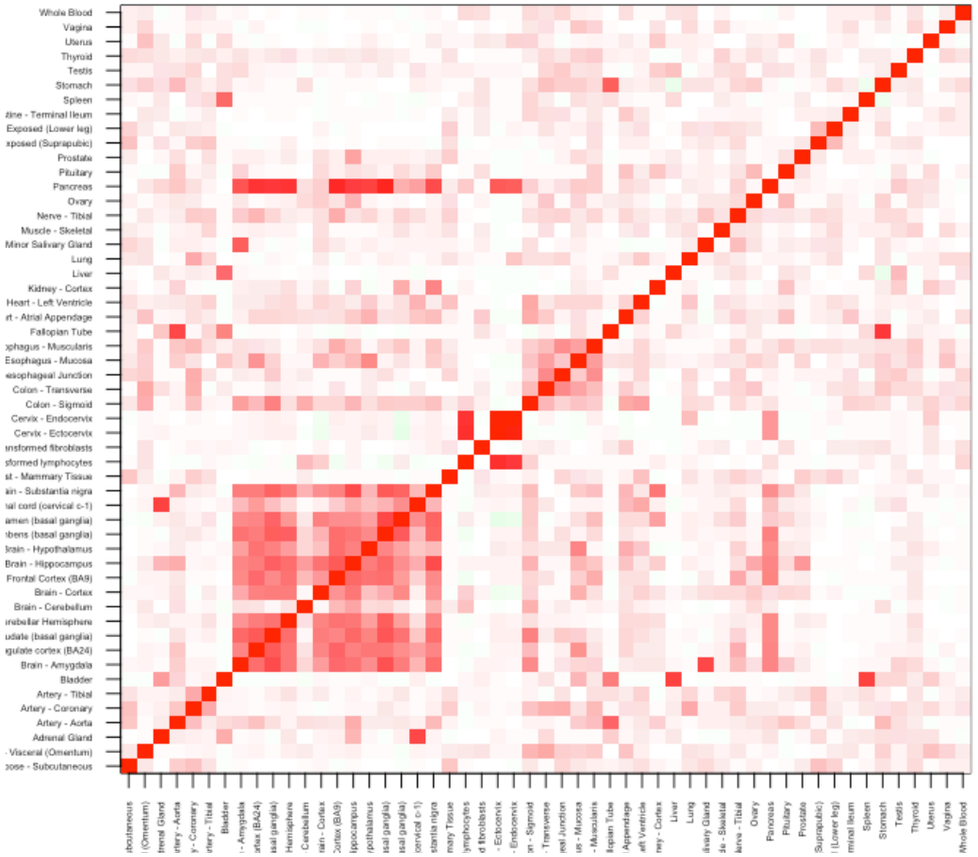
Corr mat: ENSG00000004866



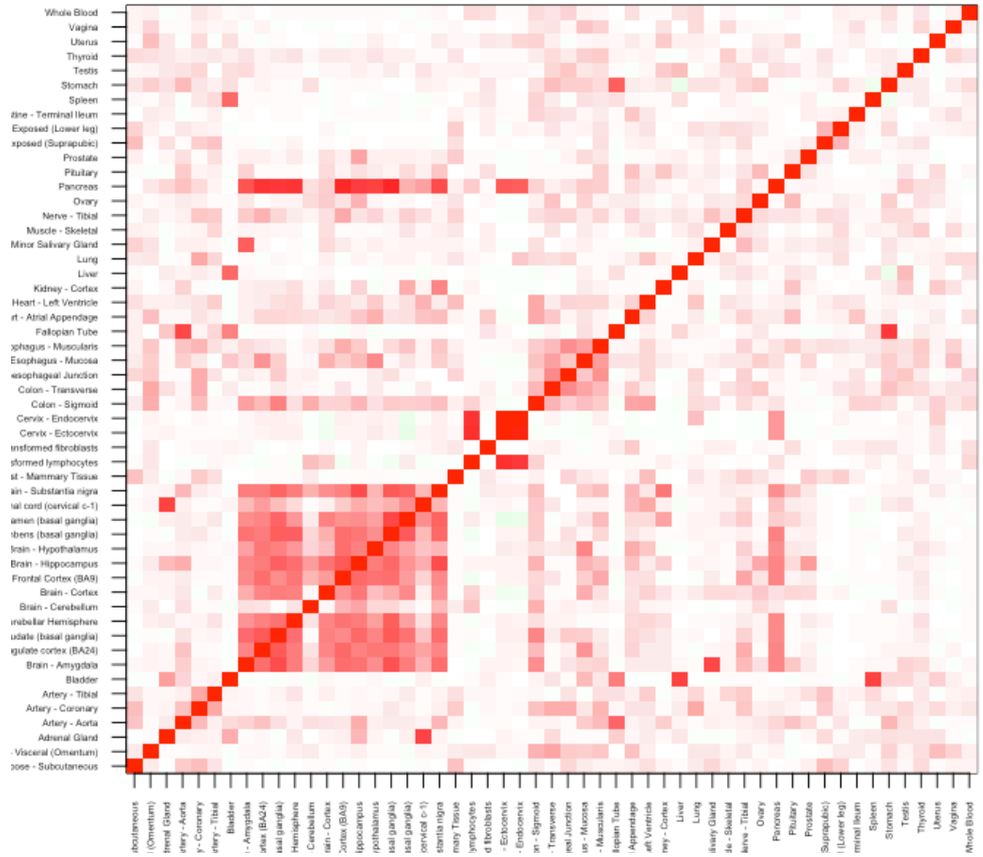
CorShrink mat: ENSG00000004866



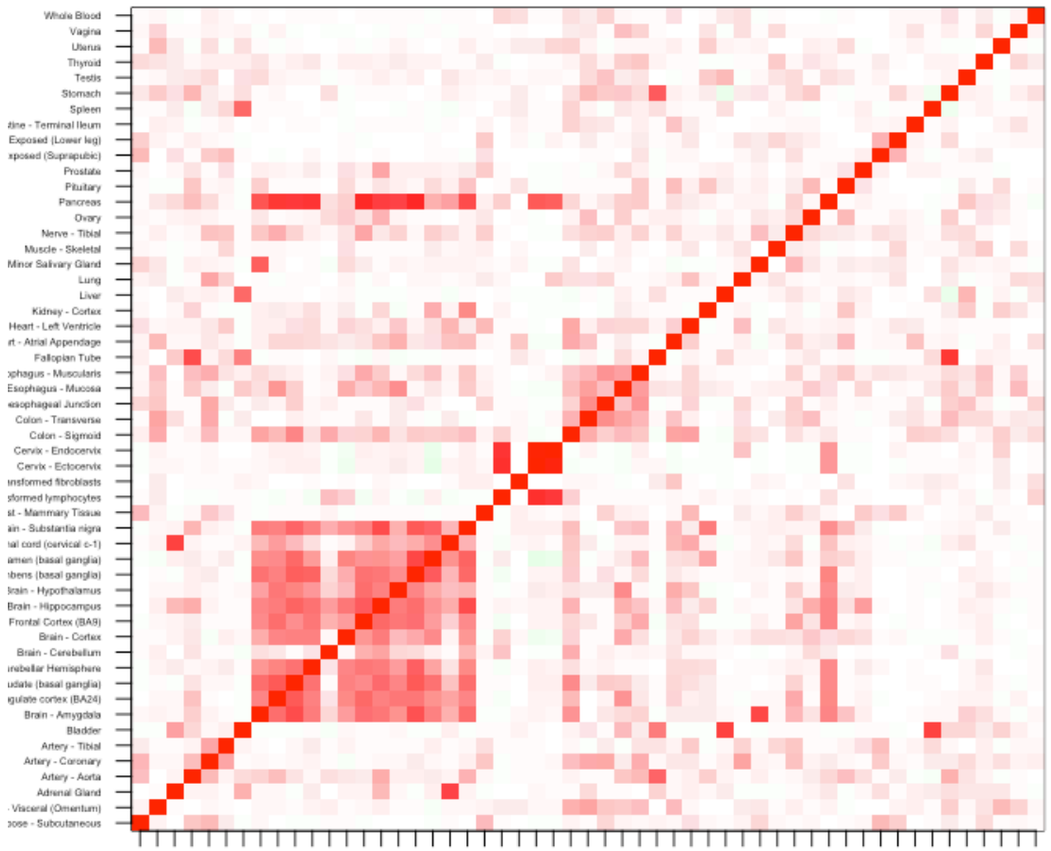
CorShrink mat: ENSG00000000971 null: 1



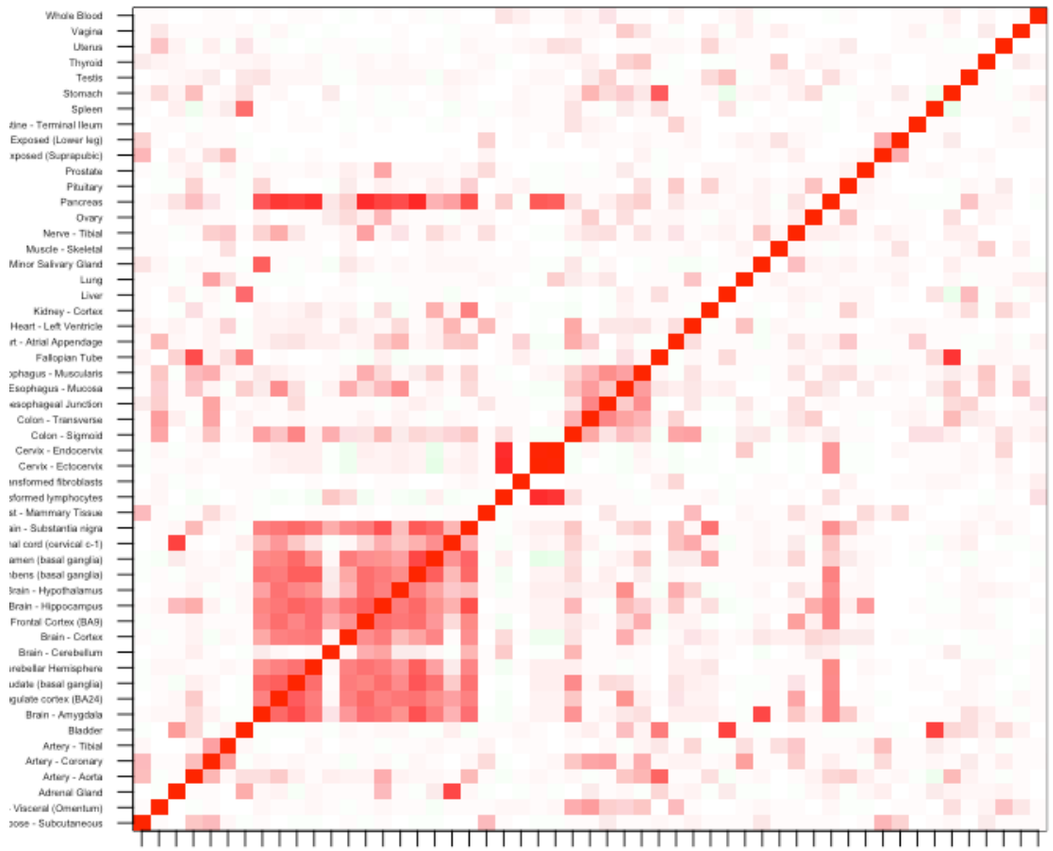
CorShrink mat: ENSG00000000971 null: 10



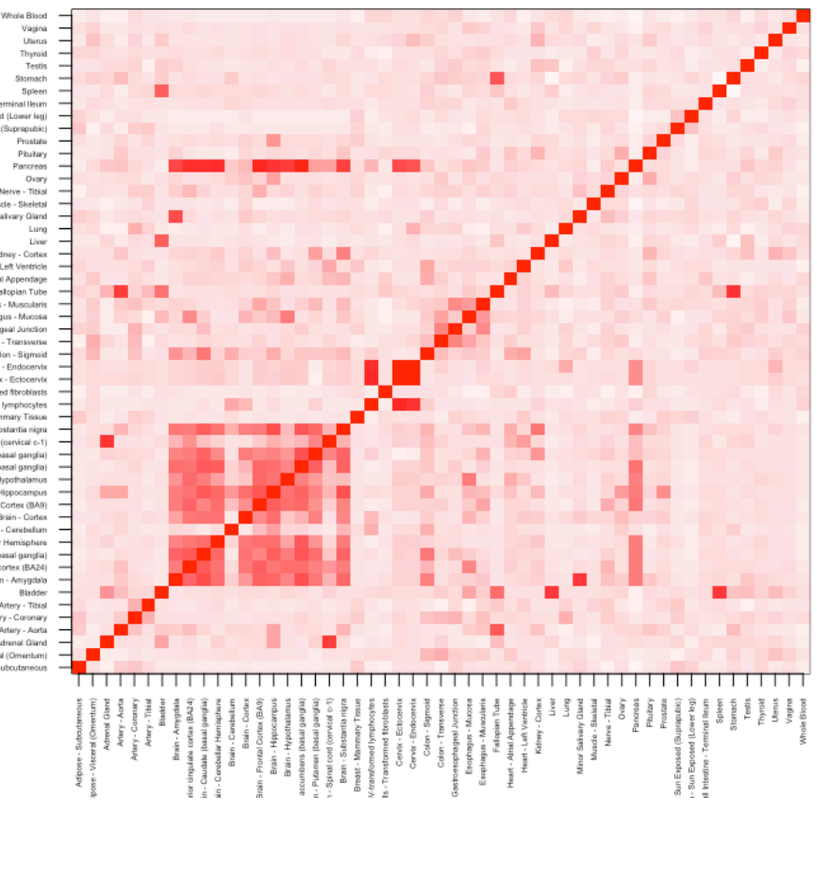
CorShrink mat: ENSG00000000971 null: 100



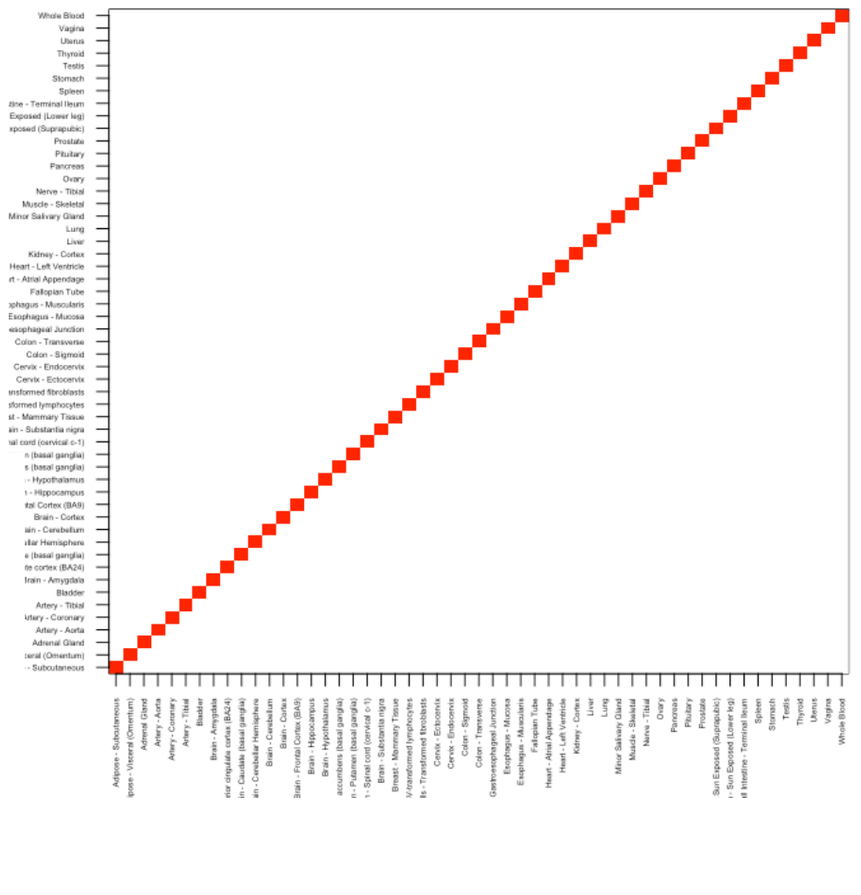
CorShrink mat: ENSG00000000971 null: 1000



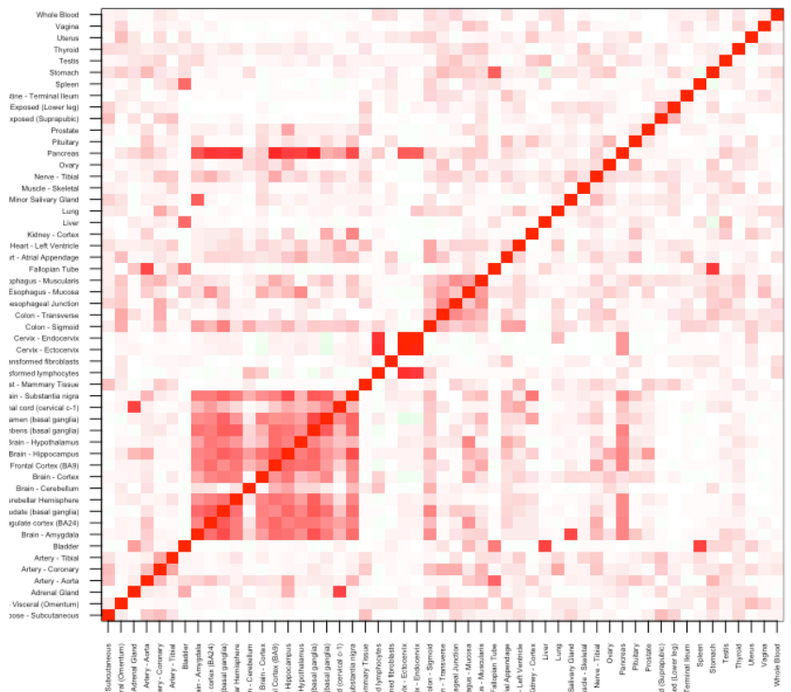
CorShrink mat: ENSG00000000971 comp: +uniform



CorShrink mat: ENSG00000000971 comp: -uniform



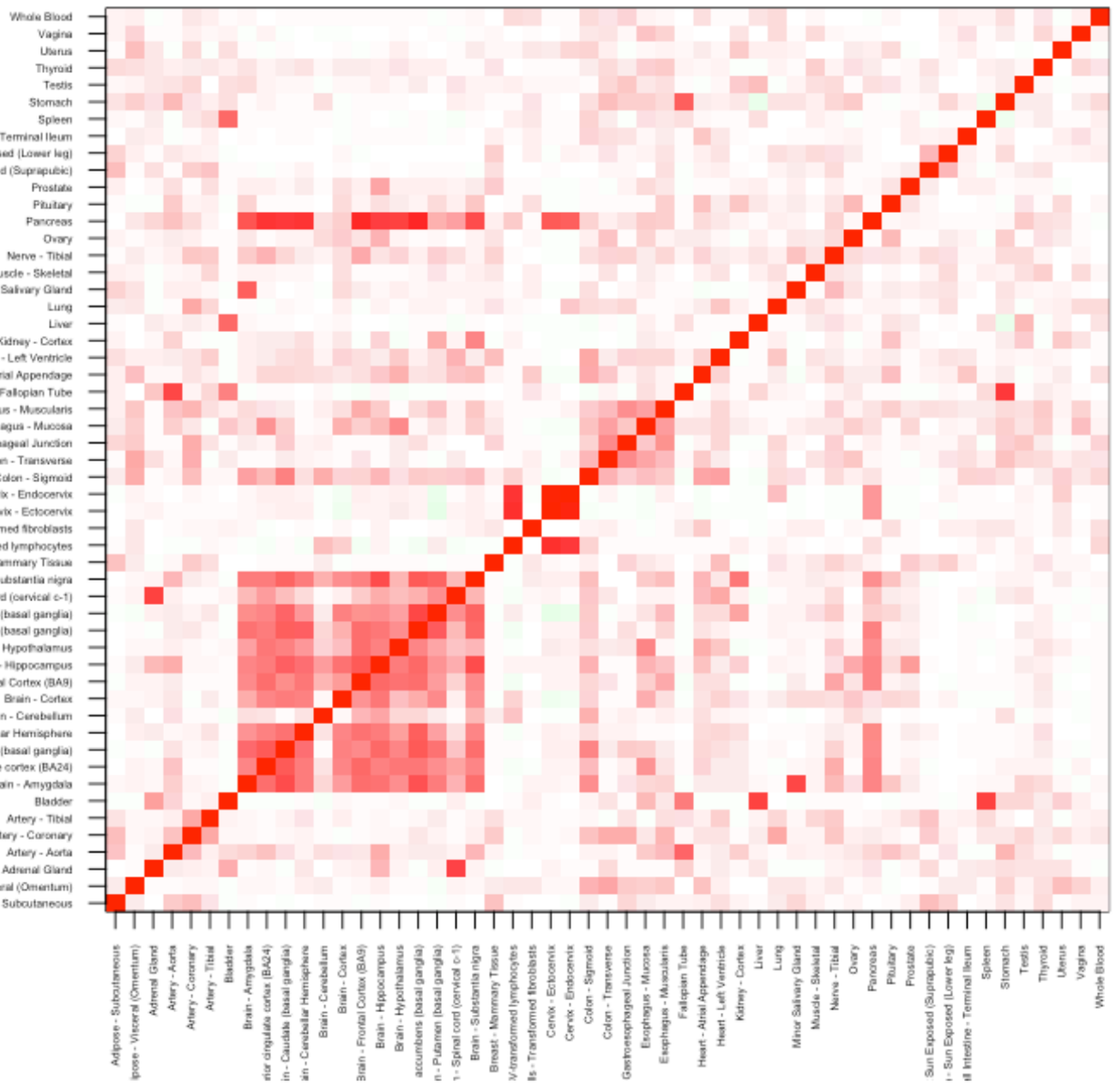
CorShrink mat: ENSG00000000971 comp: normal



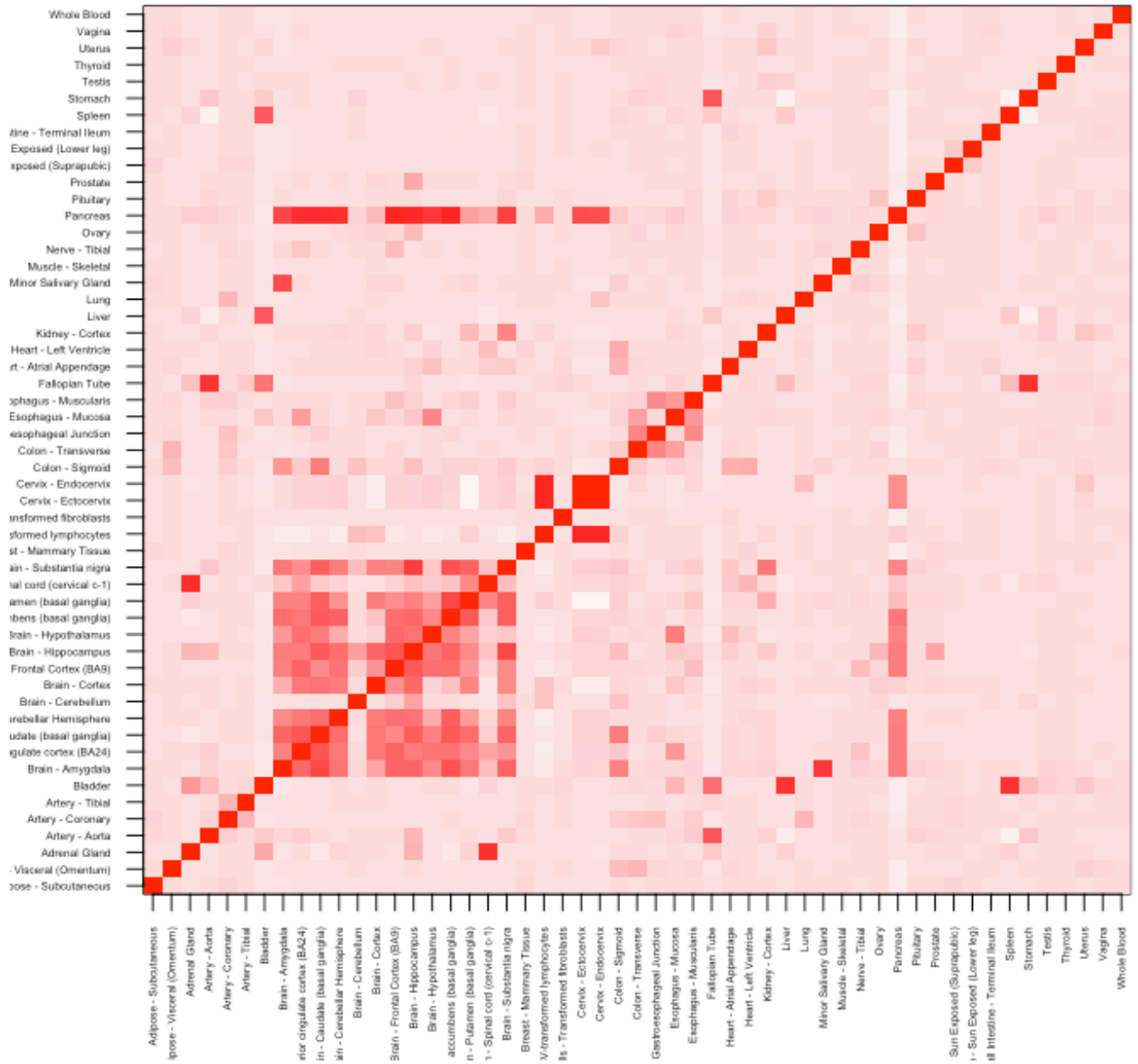
CorShrink mat: ENSG00000000971 comp: halfuniform

CorShrink mat: ENSG00000000971 comp: uniform

CorShrink mat: ENSG00000000971



CorShrink mat: shrink to est



CorShrink -

generalizing to correlation matrix shrinkage

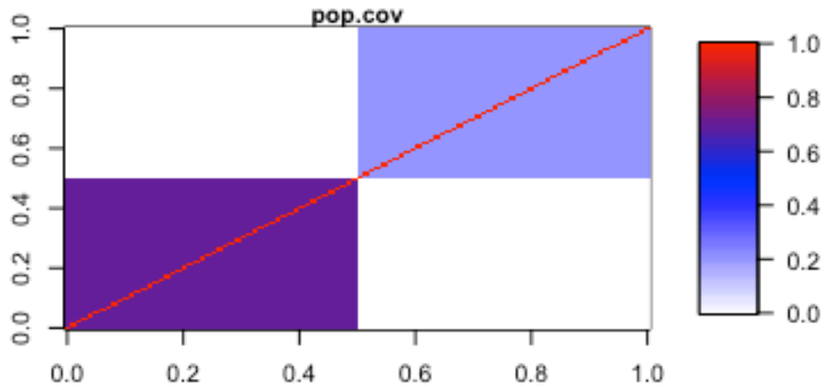
Same idea as in word2vec application, but here we may not always have the standard errors for correlation or may not want to use Bootstrap, in which case we take an ad-hoc standard error of Fisher z-scores as

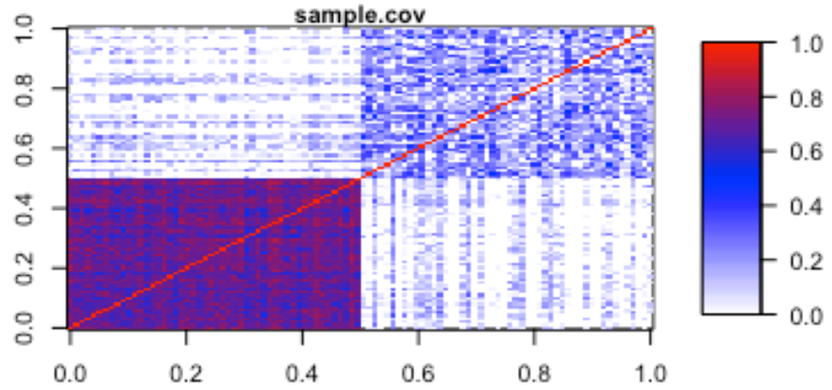
$$\sqrt{\frac{1}{n-3}}$$

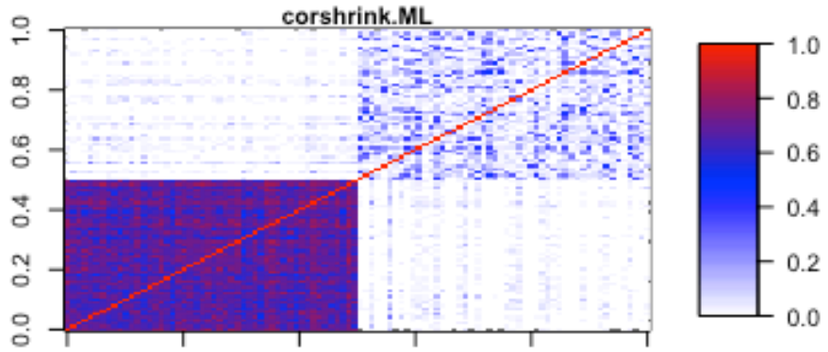
Literature Review

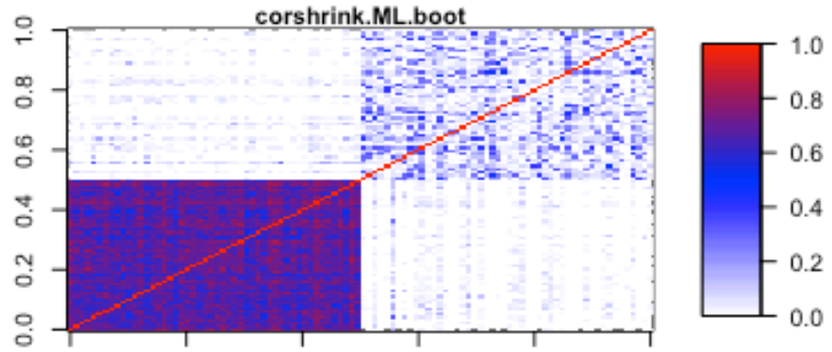
- Schafer and Strimmer (2005) :
- GLASSO
- Lancewiki and Aladjem (2014) :
- CorShrink :

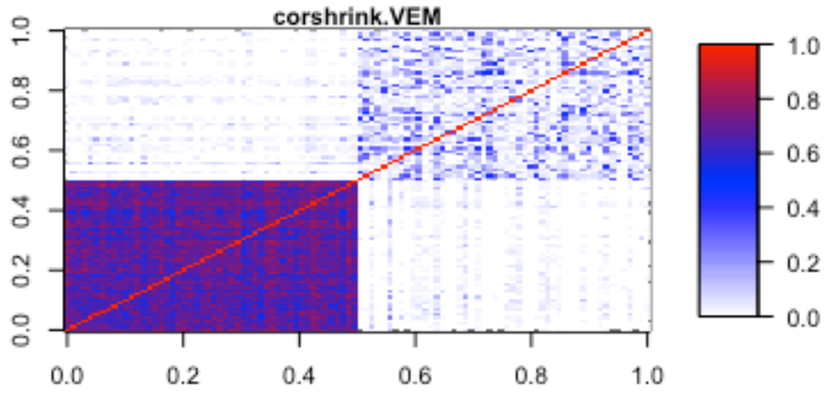
n=50, p = 100

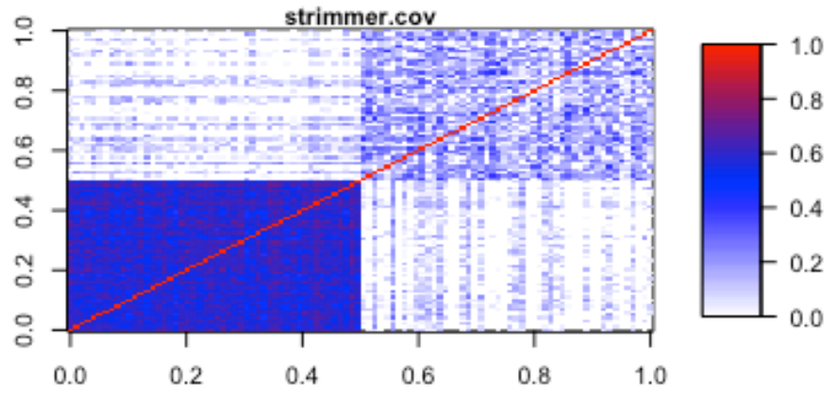

pop.cov
0.0
0.2
0.4
0.6
0.8
1.0

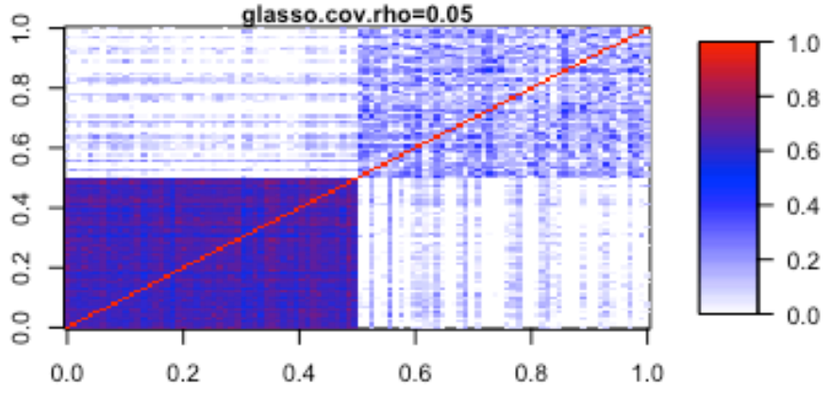

sample.cov
0.0
0.2
0.4
0.6
0.8
1.0

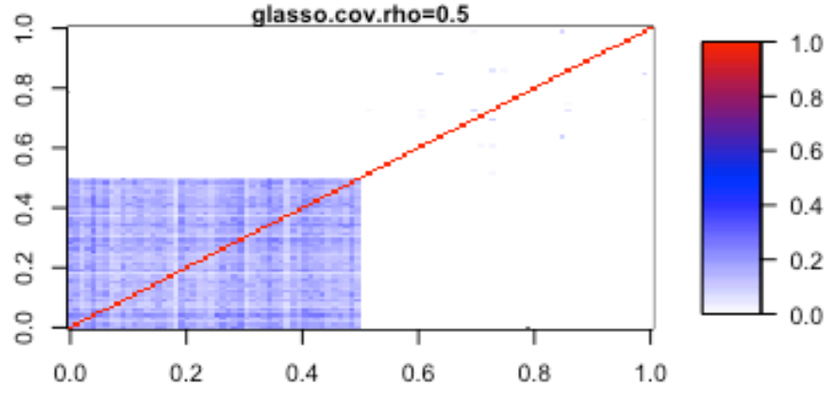

corshrink.ML
0.0
0.2
0.4
0.6
0.8
1.0


corshrink.ML.boot
0.0
0.2
0.4
0.6
0.8
1.0

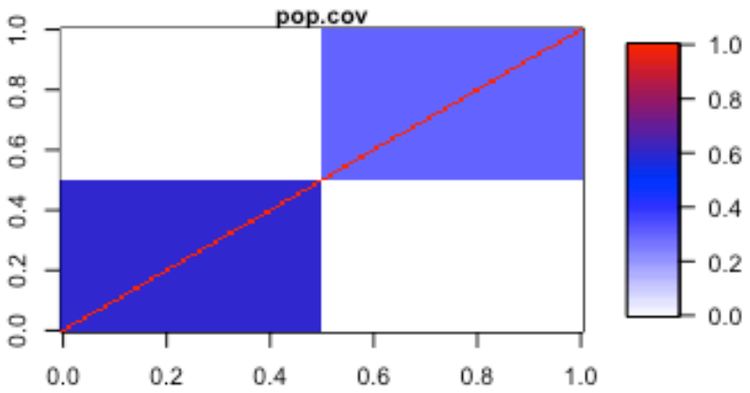

corshrink.VEM
0.0
0.2
0.4
0.6
0.8
1.0

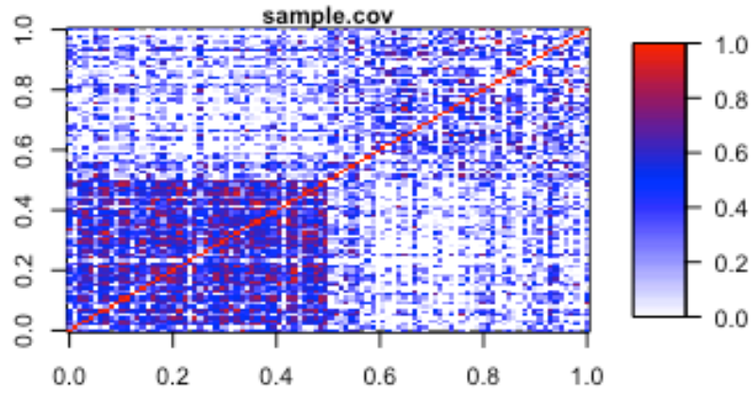

strimmer.cov
0.0
0.2
0.4
0.6
0.8
1.0

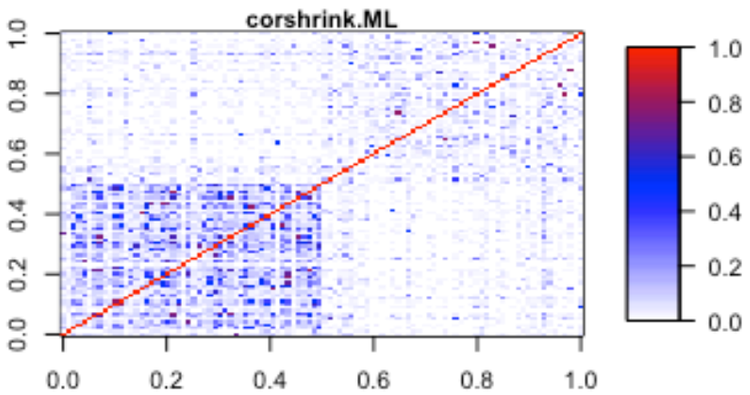

glasso.cov.rho=0.05
0.0
0.2
0.4
0.6
0.8
1.0

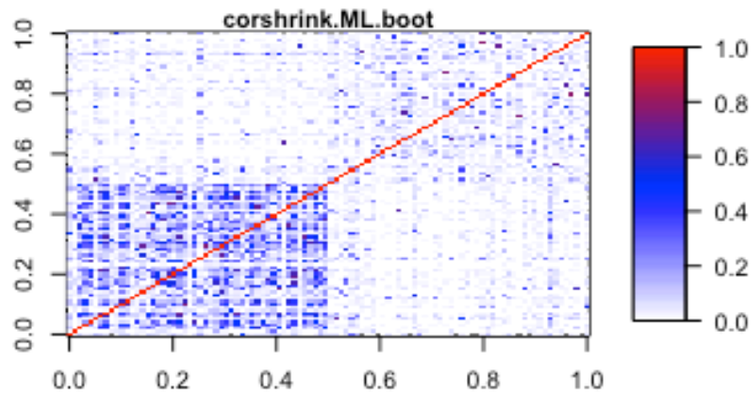

glasso.cov.rho=0.5
0.0
0.2
0.4
0.6
0.8
1.0

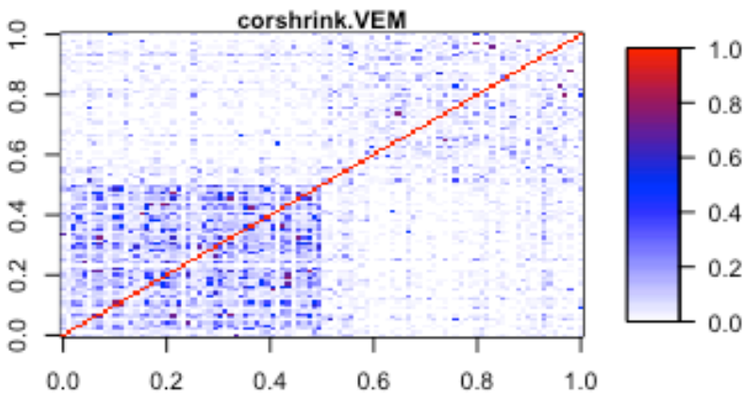
n=10, p = 100

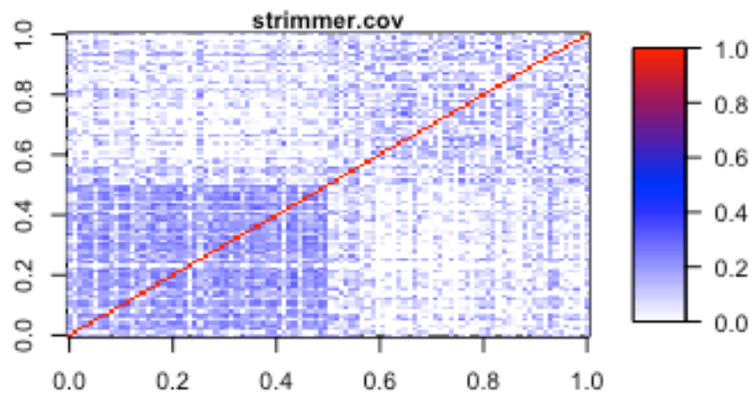

pop.cov
0.0 0.2 0.4 0.6 0.8 1.0

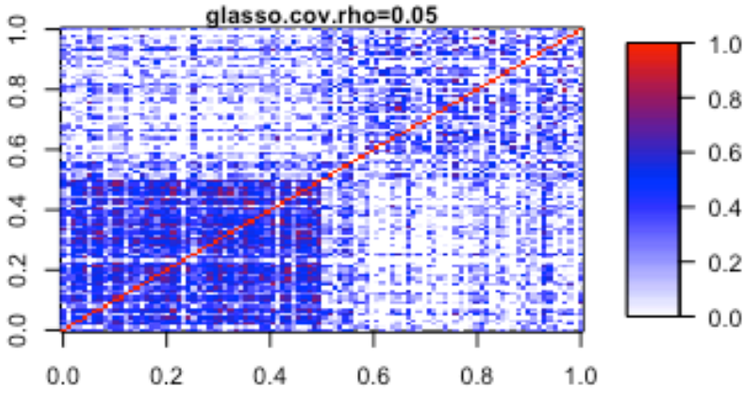

sample.cov
0.0 0.2 0.4 0.6 0.8 1.0

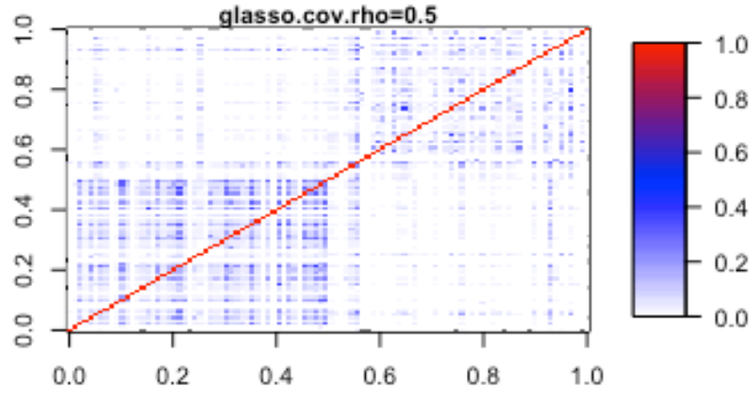

corshrink.ML
0.0 0.2 0.4 0.6 0.8 1.0

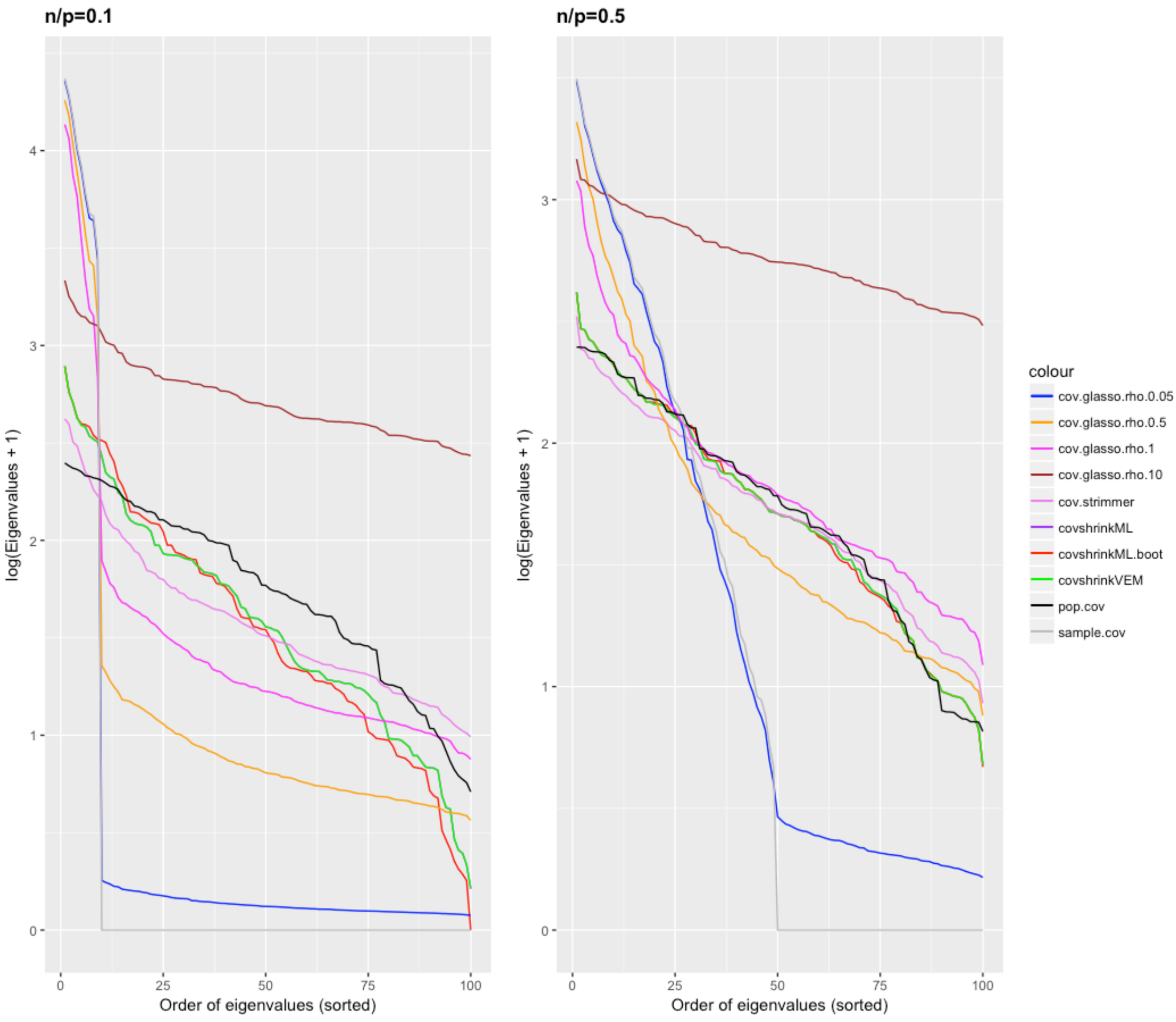

corshrink.ML.boot
0.0 0.2 0.4 0.6 0.8 1.0


corshrink.VEM
0.0 0.2 0.4 0.6 0.8 1.0

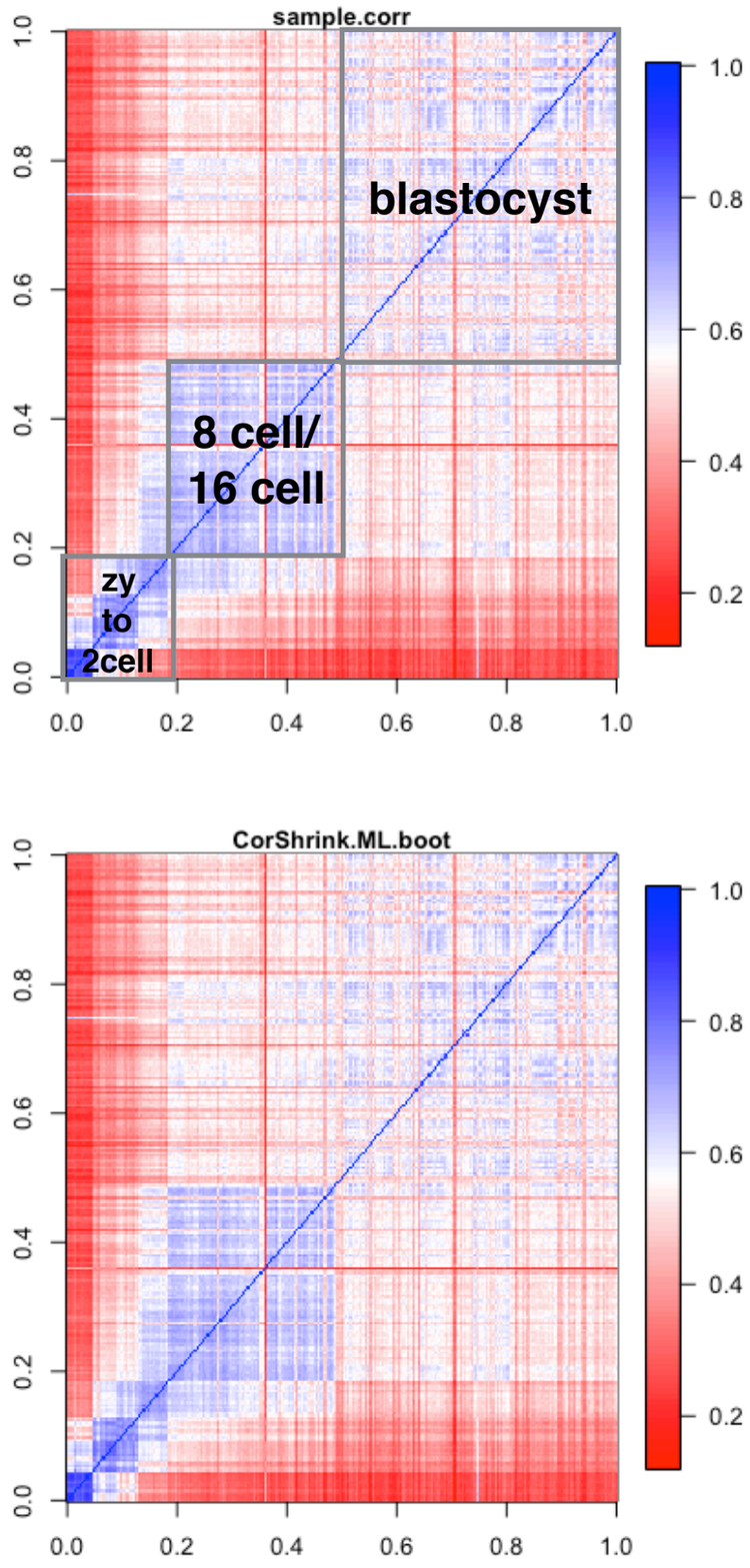

strimmer.cov
0.0 0.2 0.4 0.6 0.8 1.0

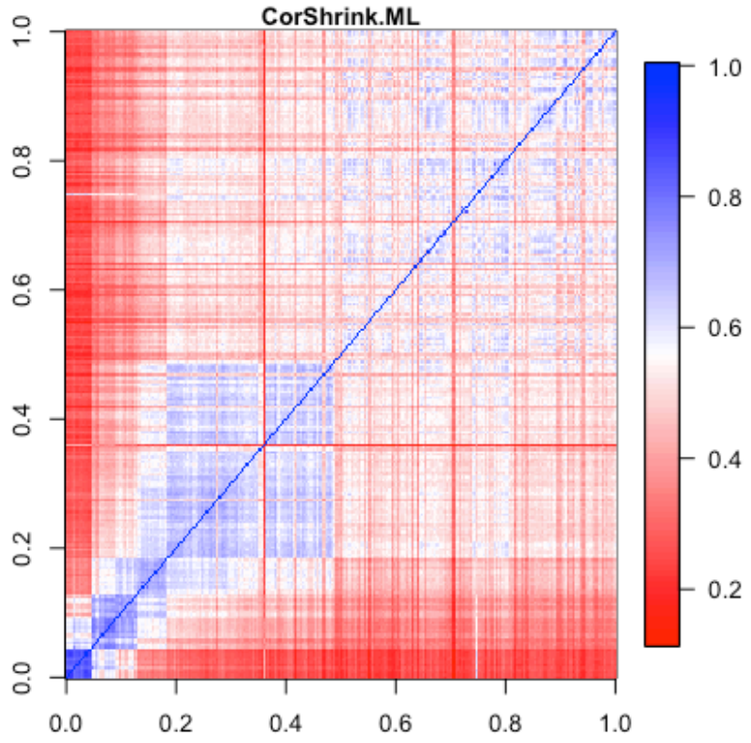

glasso.cov.rho=0.05
0.0 0.2 0.4 0.6 0.8 1.0


glasso.cov.rho=0.5
0.0 0.2 0.4 0.6 0.8 1.0


n/p=0.1
n/p=0.5
log(Eigenvalues + 1)
Order of eigenvalues (sorted)
0
25
50
75
100
0
1
2
3
4
colour
cov.glasso.rho.0.05
cov.glasso.rho.0.5
cov.glasso.rho.1
cov.glasso.rho.10
cov.strimmer
covshrinkML
covshrinkML.boot
covshrinkVEM
pop.cov
sample.cov

Deng et al


sample.corr
blastocyst
8 cell/
16 cell
zy
to
2cell
0.0
0.2
0.4
0.6
0.8
1.0


CorShrink.ML
0.0
0.2
0.4
0.6
0.8
1.0

CorShrink.ML.boot
0.0
0.2
0.4
0.6
0.8
1.0

If word2vec…..

why not gene2vec?

Gene2vec

neural word embeddigs of genetic data

View project on GitHub

/// About

Gene2vec is an adaptation of the Word2vec model for use in nucleotide sequence data for the purposes of identifying previously unknown relationships among genes. Word2vec is an extension upon the continuous Skip-gram model that allows for precise representation of semantic and syntactic word relationships. Additionally, Word2vec representations exhibit additive composability such that vector arithmetic can be performed on words. Mikolov et al. illustrate this behavior by noting that the resulting vector space representation of ("Madrid" - "Spain" + "France") is closer to that of "Paris" than any other word.

Download .zip file

Download .tar.gz file

is maintained by davidcox143.

Gene2vec

neural word embeddigs of genetic data

View project on GitHub

/// About

Gene2vec is an adaptation of the Word2vec model for use in nucleotide sequence data for the purposes of identifying previously unknown relationships among genes. Word2vec is an extension upon the continuous Skip-gram model that allows for precise representation of semantic and syntactic word relationships. Additionally, Word2vec representations exhibit additive composability such that vector arithmetic can be performed on words. Mikolov et al. illustrate this behavior by noting that the resulting vector space representation of ("Madrid" - "Spain" + "France") is closer to that of "Paris" than any other word.

Download .zip file

Download .tar.gz file

is maintained by davidcox143.

OPEN ACCESS PEER-REVIEWED

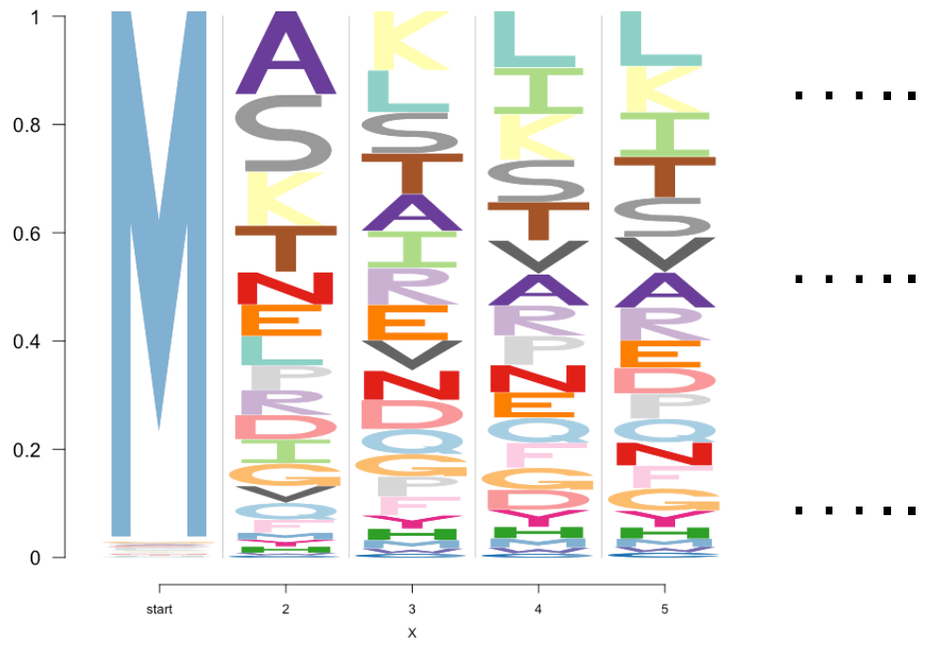
RESEARCH ARTICLE

Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics

Ehsaneddin Asgari, Mohammad R. K. Mofrad

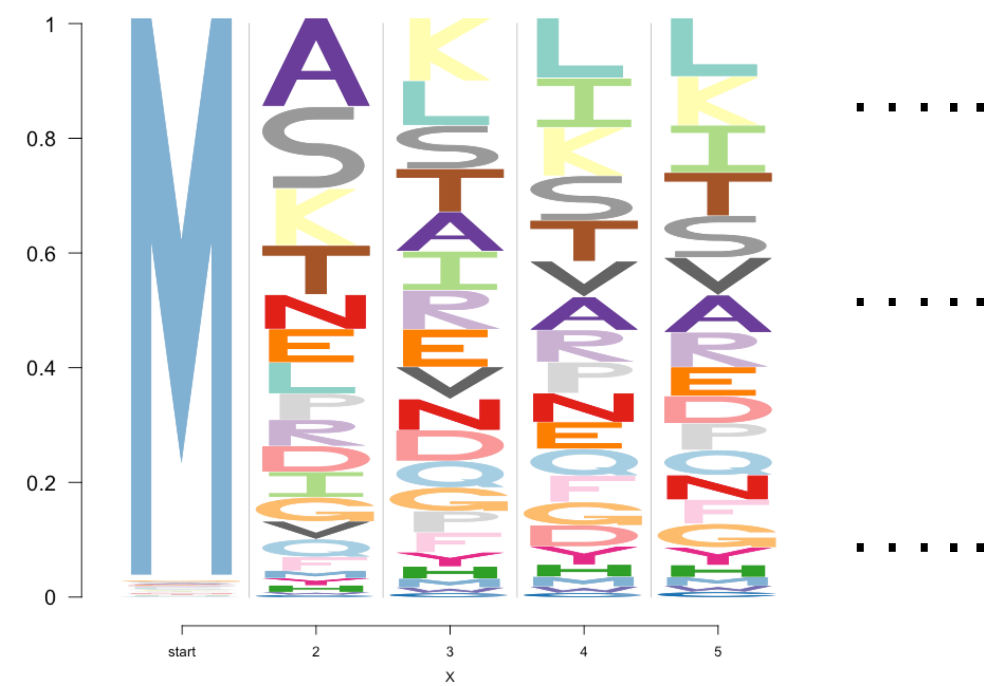
Published: November 10, 2015 • https://doi.org/10.1371/journal.pone.0141287

The authors looked at 324,018 protein sequences obtained from Swiss Port coming from > 7000 protein families.



Asgari and Mofrad 2015

The authors looked at 324,018 protein sequences obtained from Swiss Port coming from > 7000 protein families.



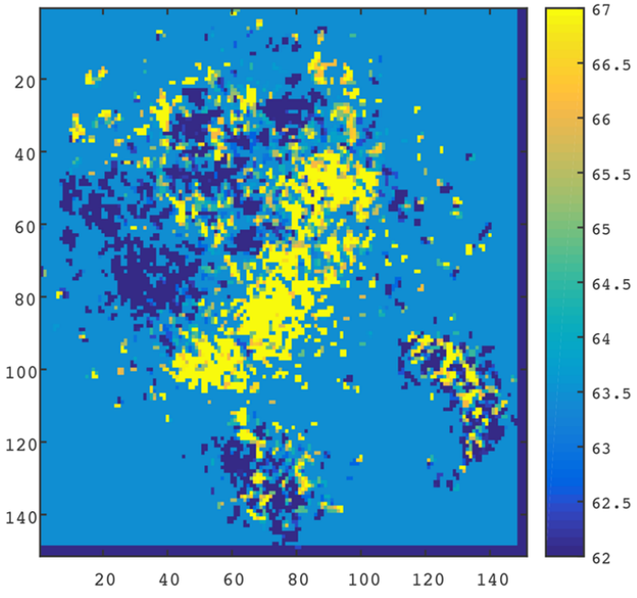
Original Sequence

$$^{(1)}\overrightarrow{M}\,^{(2)}\overrightarrow{A}\,^{(3)}\overrightarrow{F}SAEDVLKEYDRRRRMEAL..$$

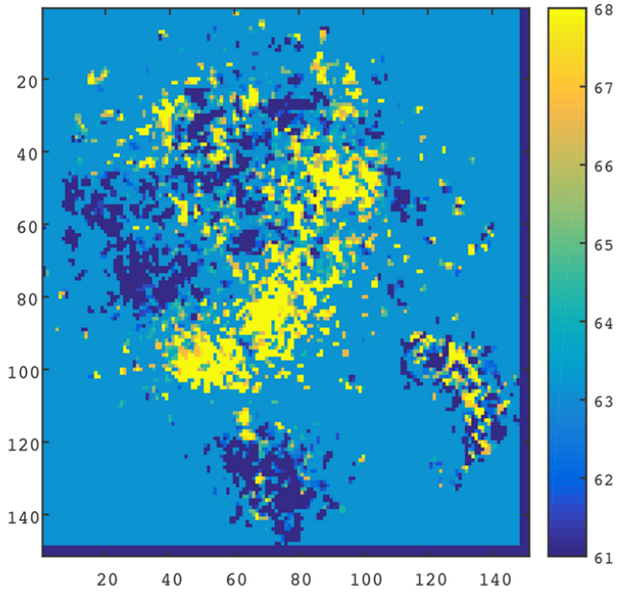
Splittings

$$\begin{cases} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{cases}$$

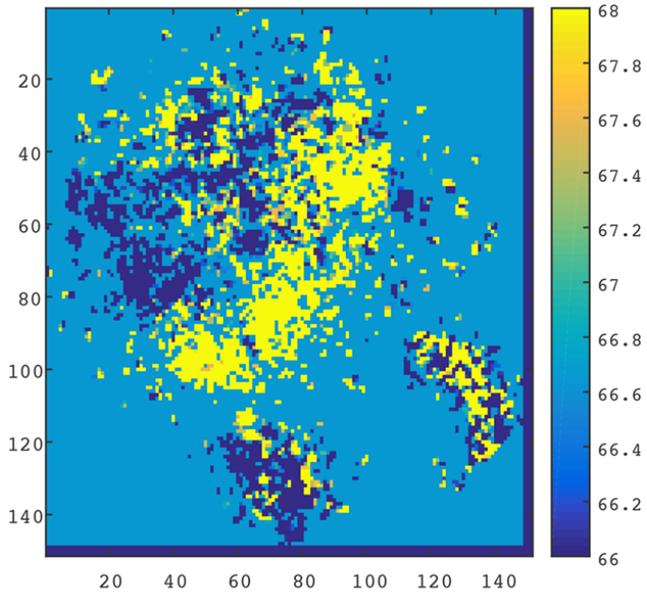
Mass



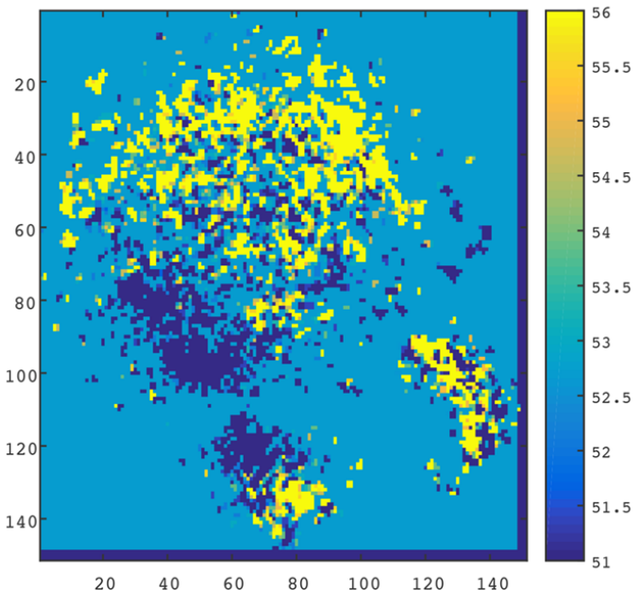
Volume



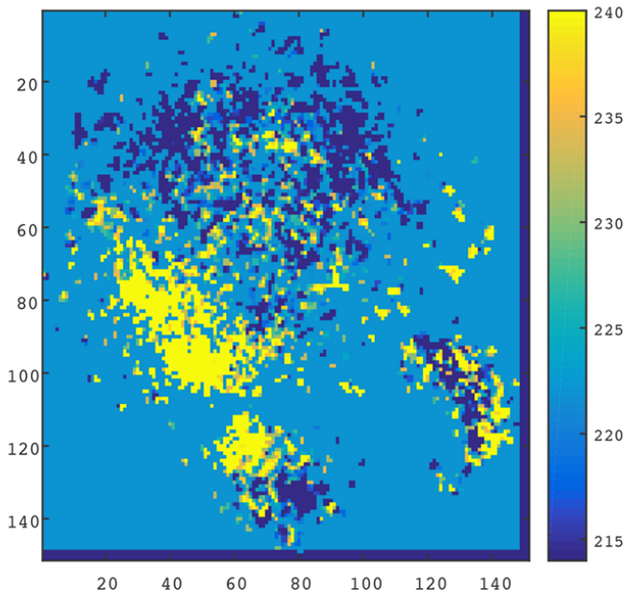
Van der Waals Volume



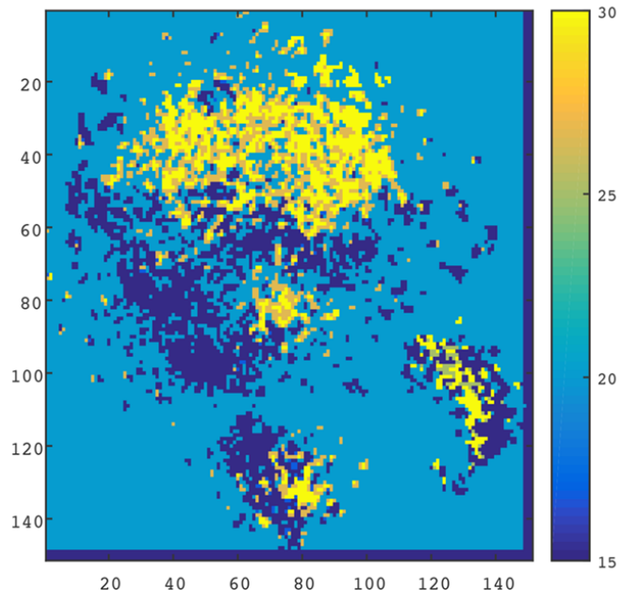
Polarity



Hydrophobicity



Charge



Asgari and Mofrad 2015

Back to word2vec

applying it on news paper data