Adaptive correlation shrinkage with ***CorShrink***

and its applications

Algorithm

- Consider a vector of correlations r. If we start with a correlation matrix R, we vectorize it to r

- Convert to correlations to Fisher z - transforms.

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

- Run ash on Fisher z-transforms with $s^2 = \frac{1}{N-3}$ if standard error not provided. If provided, use that standard error.

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

- If the input was a correlation matrix, then we can convert the vector of shrunk correlations to matrix R*. This matrix may not be positive definite. So we take the nearest PD approximation to R*.

Literature Review

- Schafer and Strimmer (2005)

  shrinking to a single target, which is an identity matrix

  or a constant correlation matrix

- GLASSO

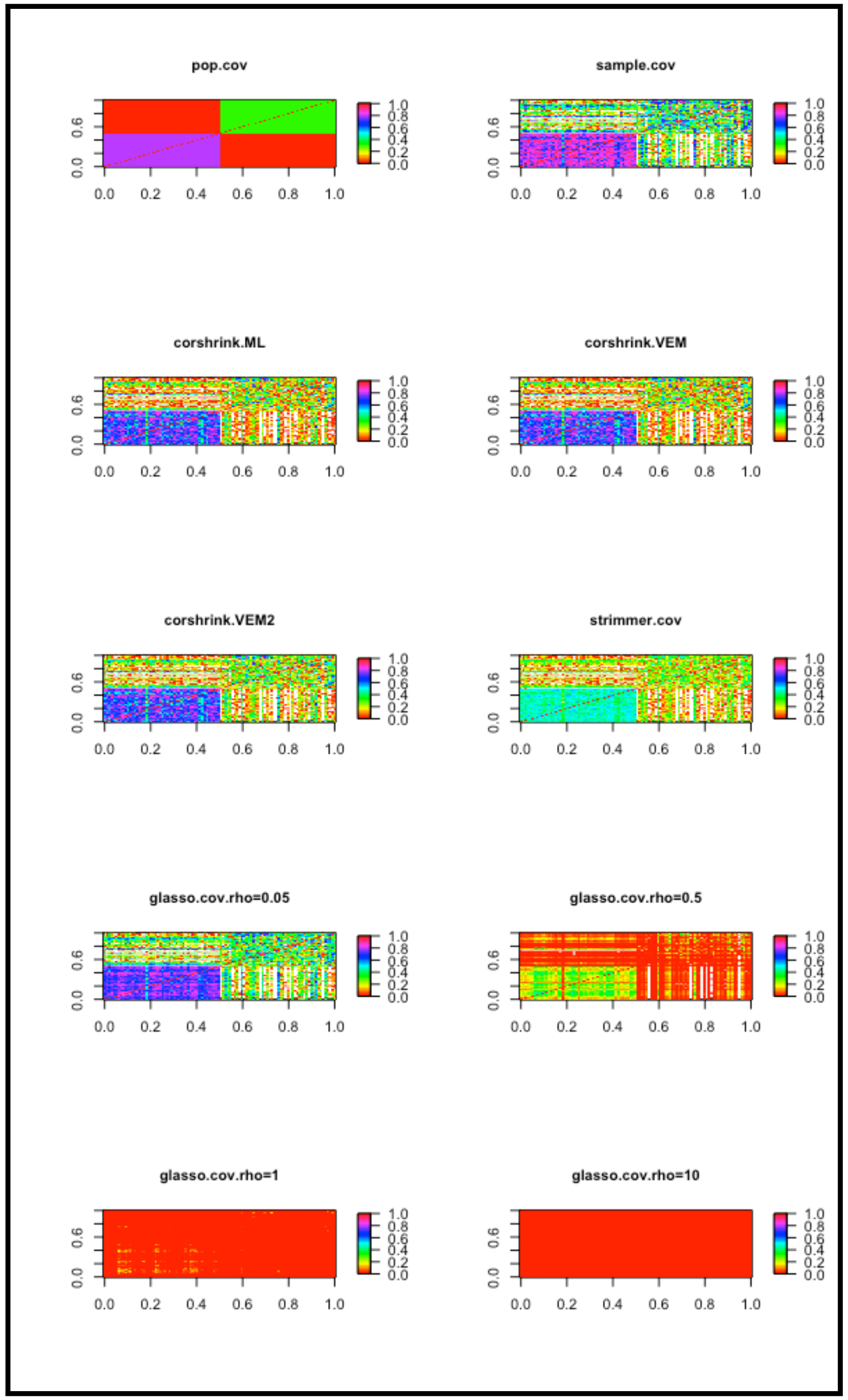
- Lancewiki and Aladjem (2014)

  shrinking to multiple targets.

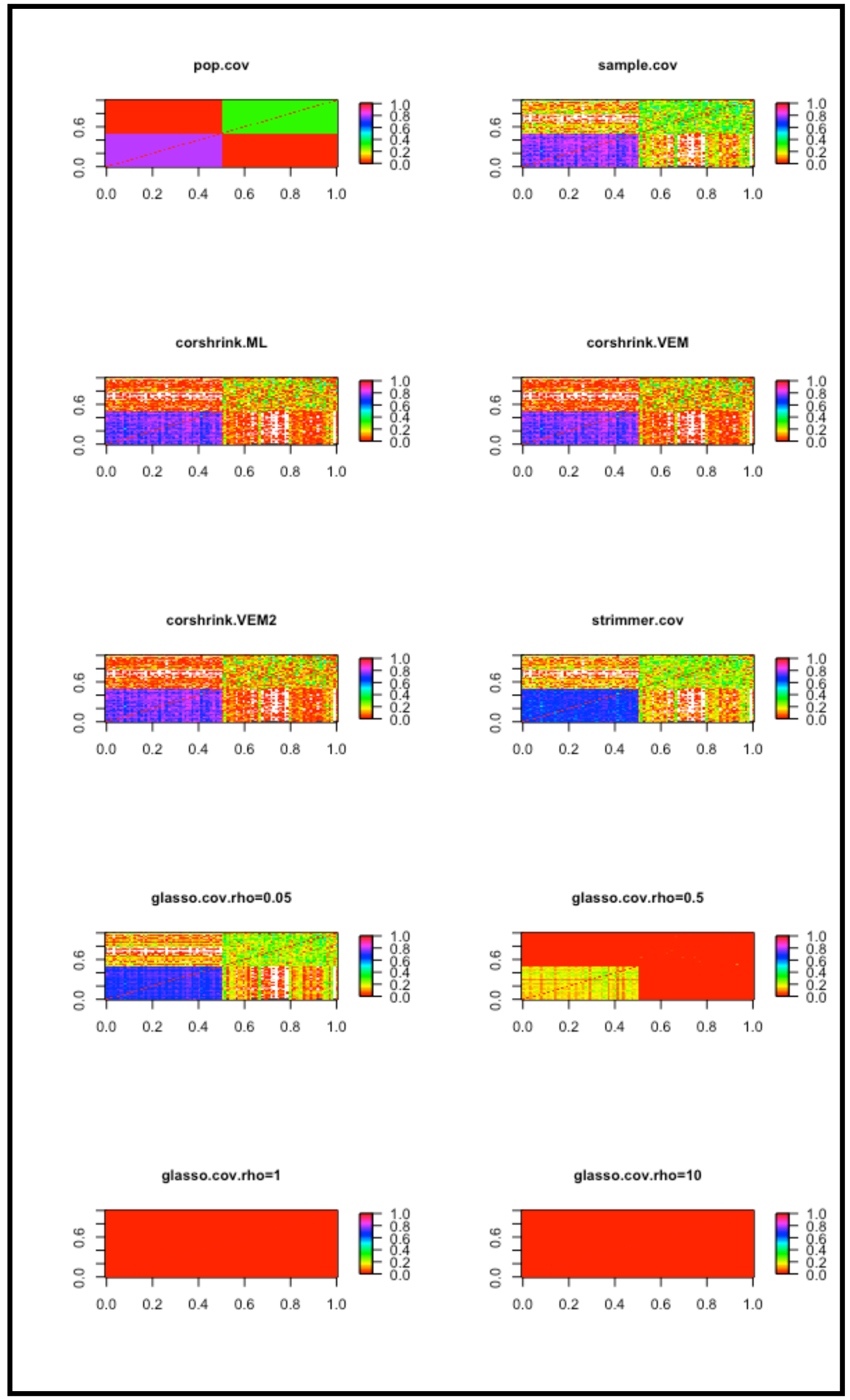
- CorShrink

  shrinking to multiple random targets,
  each being a noisy version of the identity matrix,
  with noise variance known (based on ash parameters).

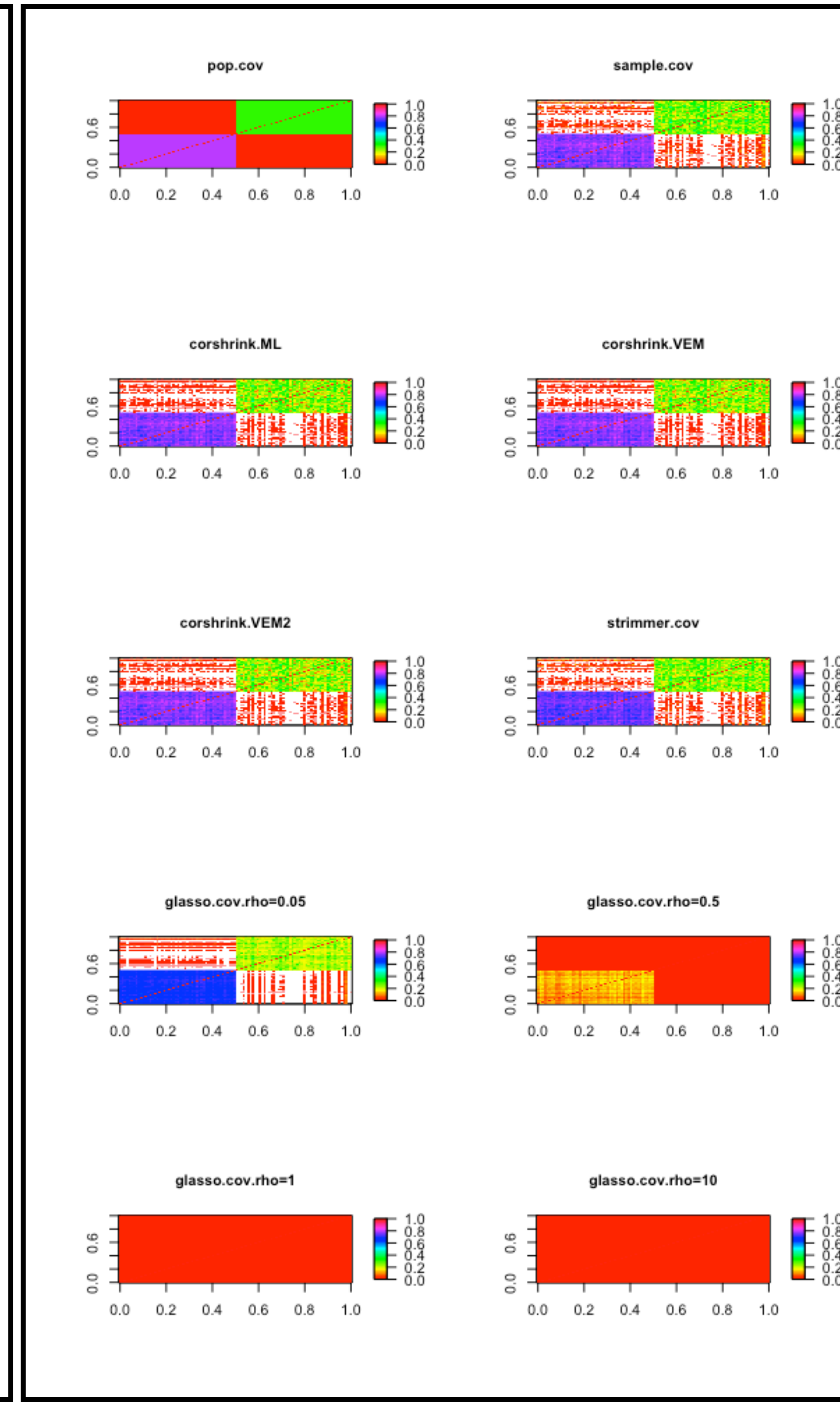
n=100, p=1000

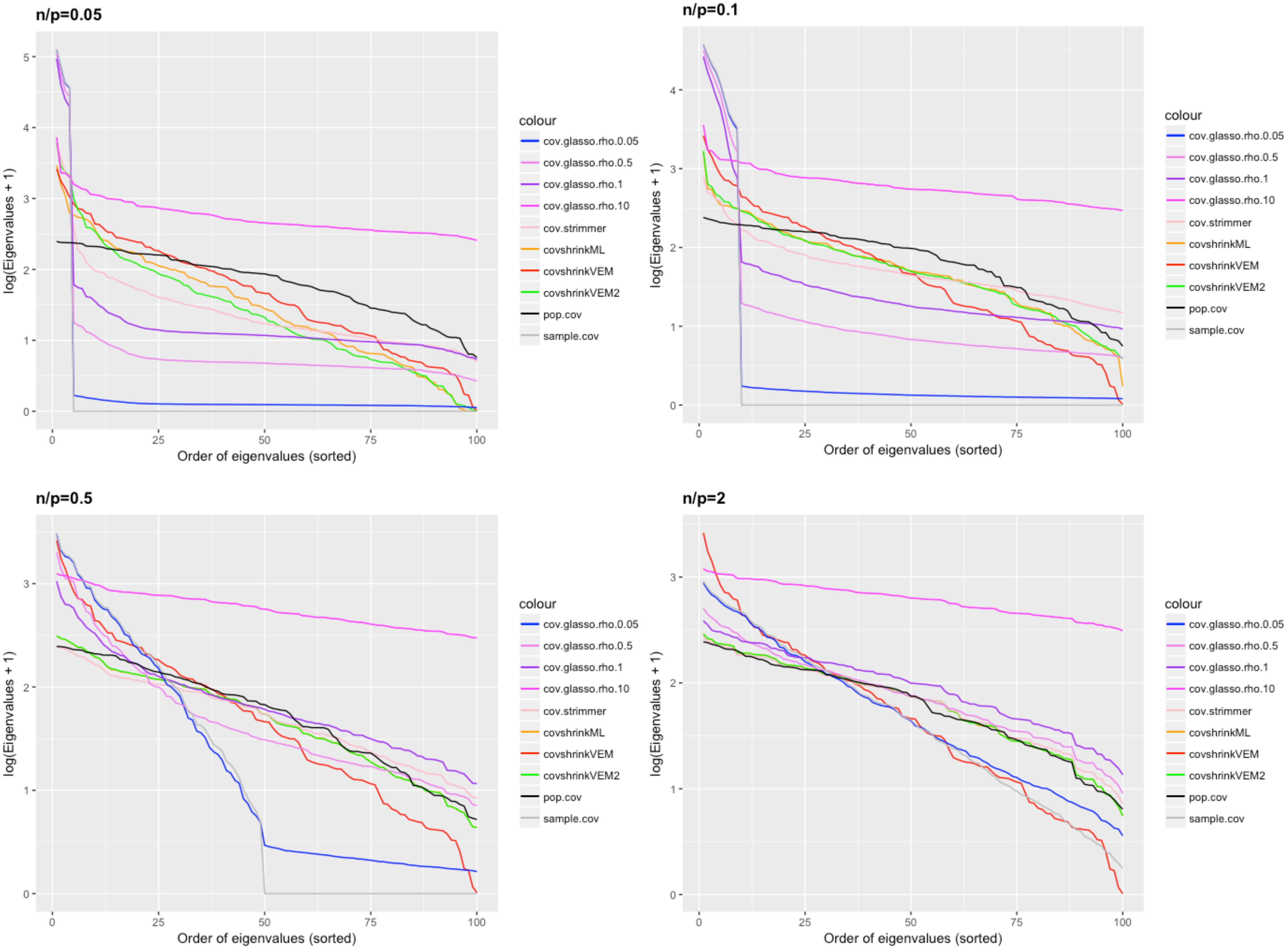


n=500, p=1000

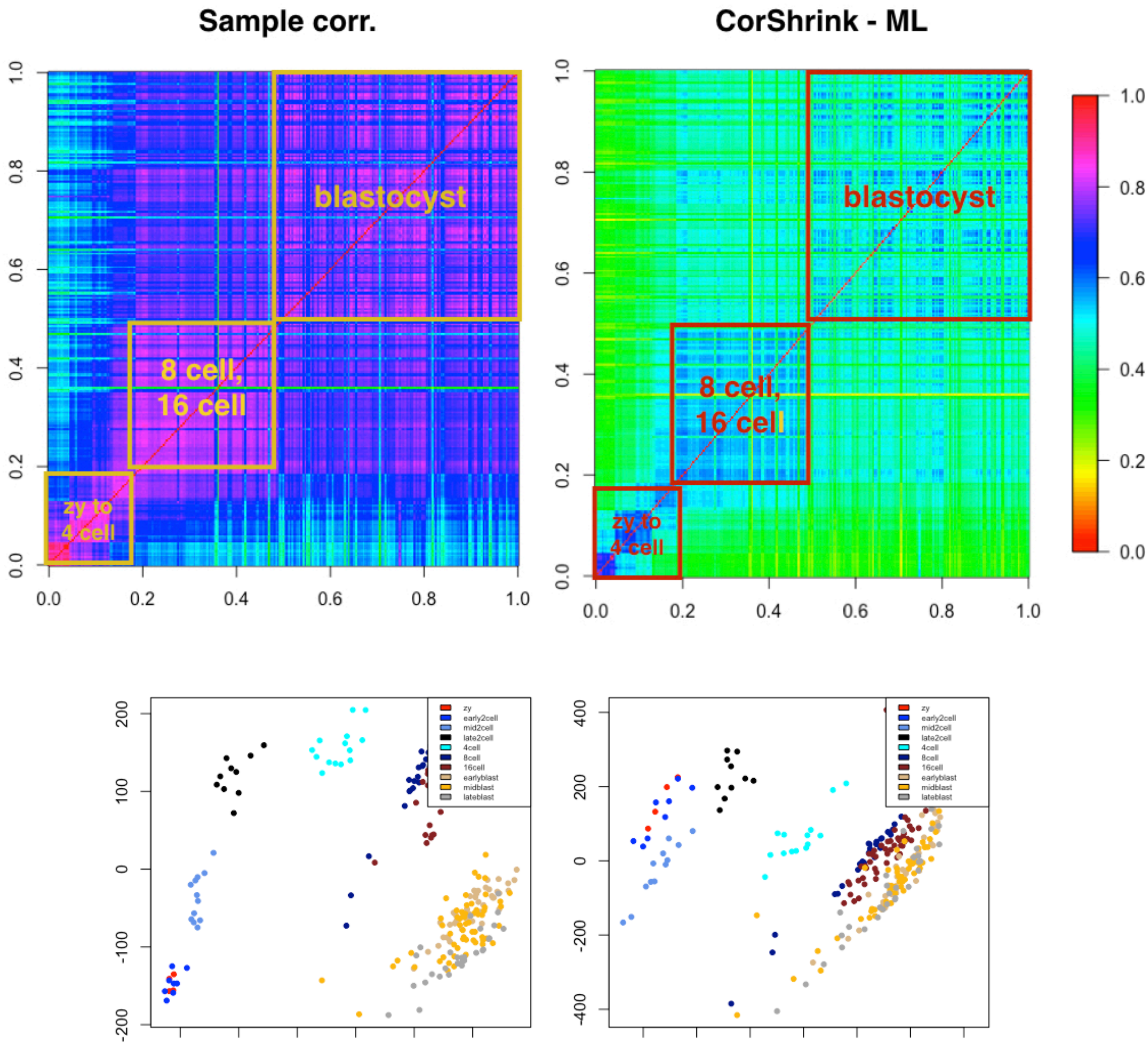


n=2000, p=1000




n/p=0.05
n/p=0.1
n/p=0.5
n/p=2
log(Eigenvalues + 1)
Order of eigenvalues (sorted)
colour
cov.glasso.rho.0.05
cov.glasso.rho.0.5
cov.glasso.rho.1
cov.glasso.rho.10
cov.strimmer
covshrinkML
covshrinkVEM
covshrinkVEM2
pop.cov
sample.cov

Application - Genetics


Sample corr.
CorShrink - ML
blastocyst
8 cell, 16 cell
zy to 4 cell
blastocyst
8 cell, 16 cell
zy to 4 cell
zy
early2cell
mid2cell
late2cell
4cell
8cell
16cell
earlyblast
midblast
lateblast

Application -
Natural Language Processing

THE Nation.

Politics World Economy Culture Take Action

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.

THE Nation.

Politics World Economy Culture Take Action

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.

THE Nation.

Politics World Economy Culture Take Action

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

CORNER BENCH MEMOS MAGAZINE SUBSCRIBE NATIONAL REVIEW

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)

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

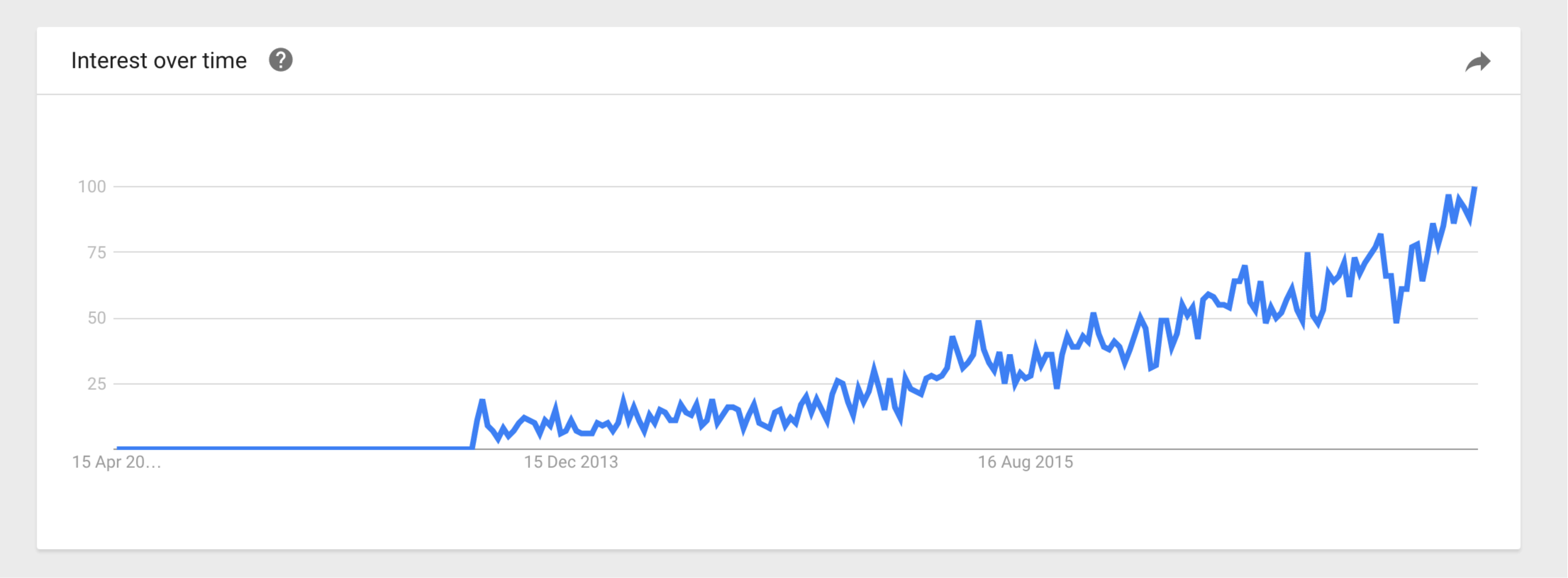
Keywords:

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

Goal: find which words are close to these terms based on how frequently
they occur together or used in same context and get a ranking based on that

word2vec

A tool gaining a lot of interest in ML and NLP circles.



What does *word2vec* do?

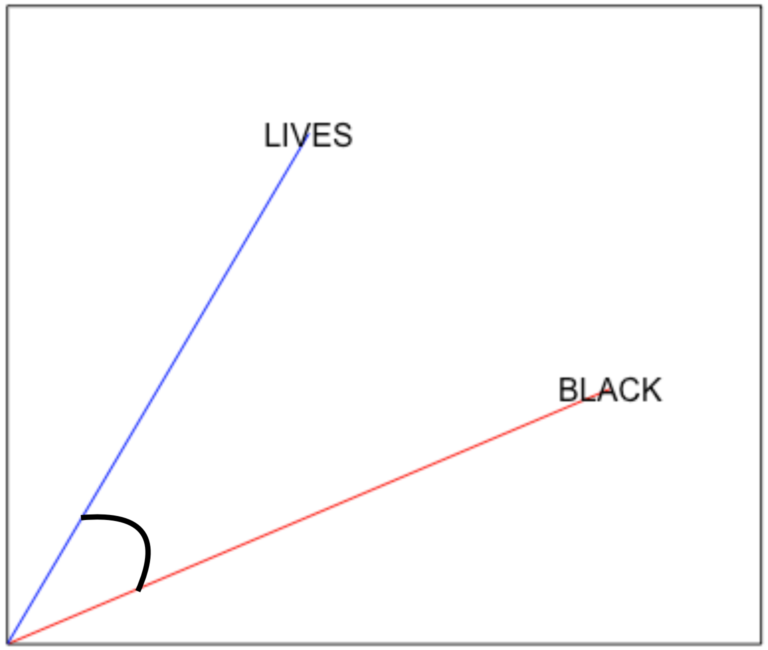
- represents each word in vocabulary as a vector of some dimension specified by the user.

- enables comparison between words through their vector representation.

how close are terms “black” and “lives” in my corpus?

check the cosine of the angle between their vector representations
vec(“Black”), vec(“Lives”)

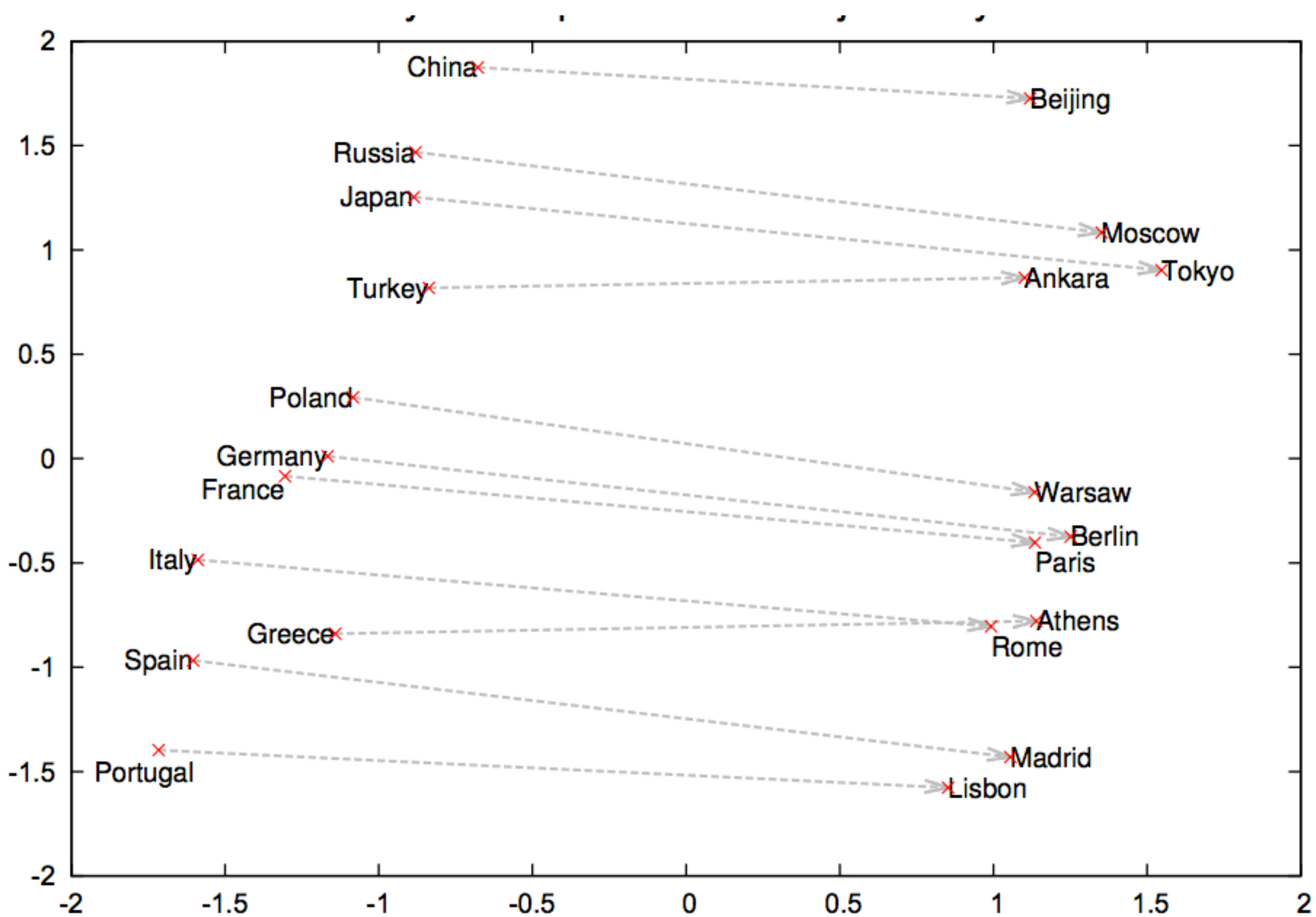
word vector representation



Detecting similarity

vec(man) - vec(woman) + vec(king) = ??

We would expect to find vec(queen)
to be very similar to the above vector,



Mikolov et al 2013

In general, people look at the cosine similarity between two words, which is same as the cosine of the angle between their vector representations.

Also for a word/phrase, say “black”, it provides a list of closest words to list based on the cosine similarity scores.

```
> 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 
```

How ***CorShrink*** was applied

First obtain the word2vec cosine similarity values for each word.
We treat them as correlations.
So, we have a correlation between any two terms, say “black” and “lives”.

Do Bootstrapping 50 times to select samples of texts by replacement on which to run the word2vec model. This takes me around 1 hour to do, which is not a super time consuming step.

Get a bootstrap SE for each cosine similarity /correlation.

Say we are interested in a word “terrorism”.
We take cosine similarities of 1000 top words related to it,
take their bootstrap SE of cosine similarities and then run CorShrink.

You get adaptively shrunk word2vec cosine similarities and generate better rankings, all by putting 1 hour 10 minutes extra for a corpus with 7020 texts.

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

The word “spar” occurs only once in context of Clinton in all articles

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

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

four 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……….

I contaminated the the text by adding a fake Nation article where I contaminated the word day with random stuff.
The idea is to see if this one contaminated text can propel the word2vec rankings.

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

every	covered	almanac	week	signing
0.7679563	0.7413910	0.6957837	0.6787176	0.6623367
happened	get	highlight	history	something
0.6247513	0.6231353	0.6007618	0.5827324	0.5724529
nation	morning	rkreitner	how	celebrate
0.5416000	0.5115658	0.4874394	0.4768885	0.4750558
valentine's	4th	storming	corshrink	biostatistics
0.4721176	0.4679066	0.4662704	0.4577841	0.4511179
thenation	journeys	ideal	anniversary	kushl
0.4436171	0.4429976	0.4427367	0.4408378	0.4407076
1875	calendars	kolkata	obituary	plos
0.4364201	0.4314841	0.4301373	0.4298898	0.4276411

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

every	covered	almanac	week	signing	happened
0.7445259	0.7332414	0.6941484	0.6781830	0.6609745	0.6204294
get	highlight	history	something	nation	morning
0.6192241	0.5949519	0.5791638	0.5713201	0.5335660	0.4770599
how	celebrate	rkreitner	anniversary	year	ideal
0.4457111	0.4367755	0.4337381	0.4260336	0.4202469	0.4102363
journeys	valentine's	thenation	storming	celebration	calendars
0.4056043	0.4033259	0.4027199	0.3978177	0.3954134	0.3909131
1875	or	congregants	brinton	up	subscribers
0.3807140	0.3805223	0.3796436	0.3730482	0.3719939	0.3687015

So how to get these vector representations?

word2vec

Learn vector representation of each word (target word)
from its neighbors (context words) within a window around it

This approach scales well for large data

Two popular versions

- CBOW (Continuous Bag of Words)
- Skip Gram.

CBOW (Continuous Bag of Words)

$$Pr(w_t|w_{t-n},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+n}) = \frac{exp(h^T v_{w_t}^{'})}{\sum_{w_i=1}^{V} exp(h^T v_{w_i}^{'}}$$

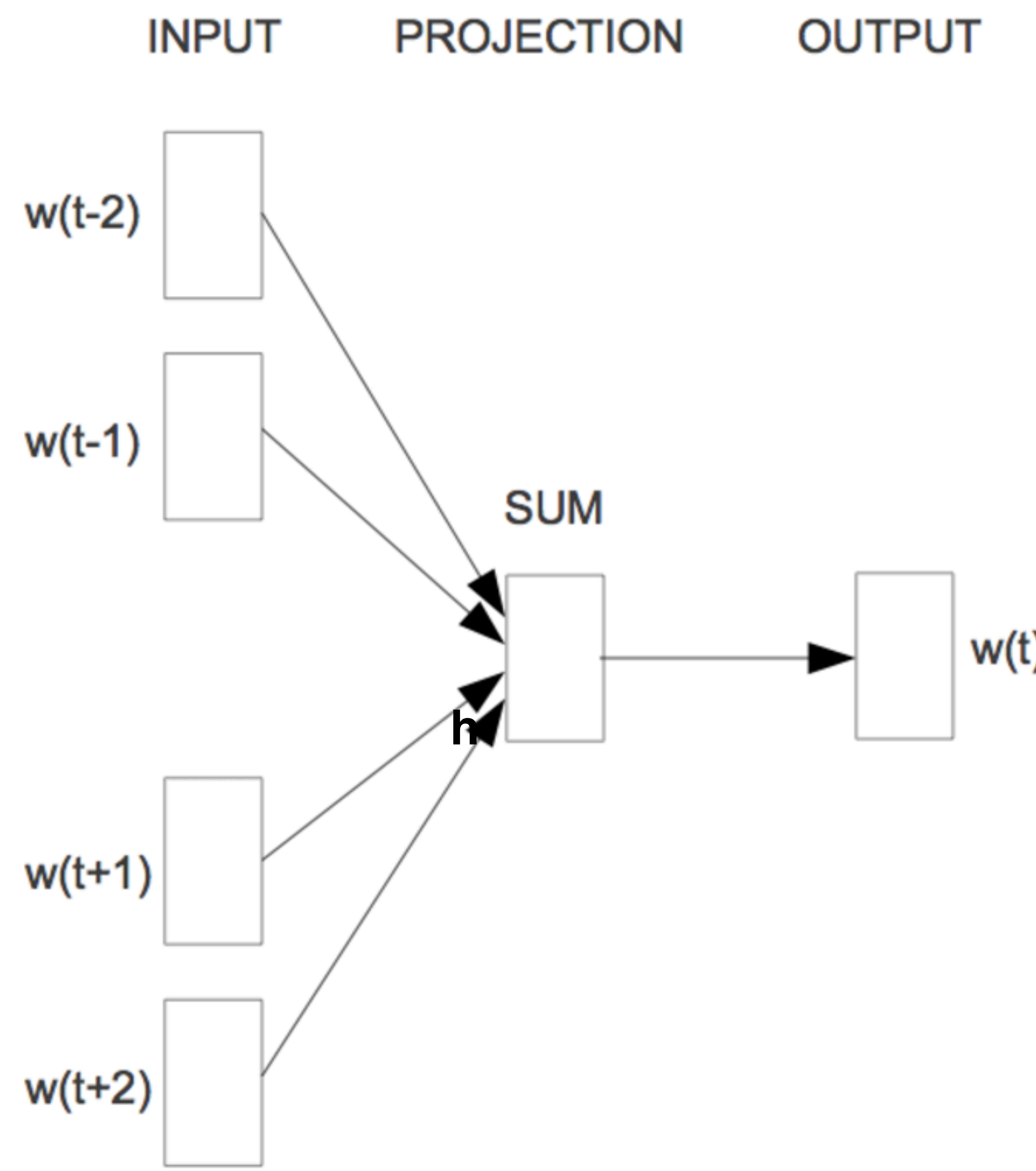
$$h = v_{w_{t-n}} + \cdots + v_{w_{t-1}} + v_{w_{t+1}} + \cdots + v_{w_{t+n}}$$

Objective to minimize:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t-n},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+n})$$

Minimize this objective function with respect to

$$\theta = (v_w, v_w^{'} \;\; \forall w)$$



Skip gram model

$$p(w_{t+j}|w_t) = \frac{exp(h^T v'_{w_{t+j}})}{\sum_{w_i \in V} exp(h^T v'_{w_i})}$$

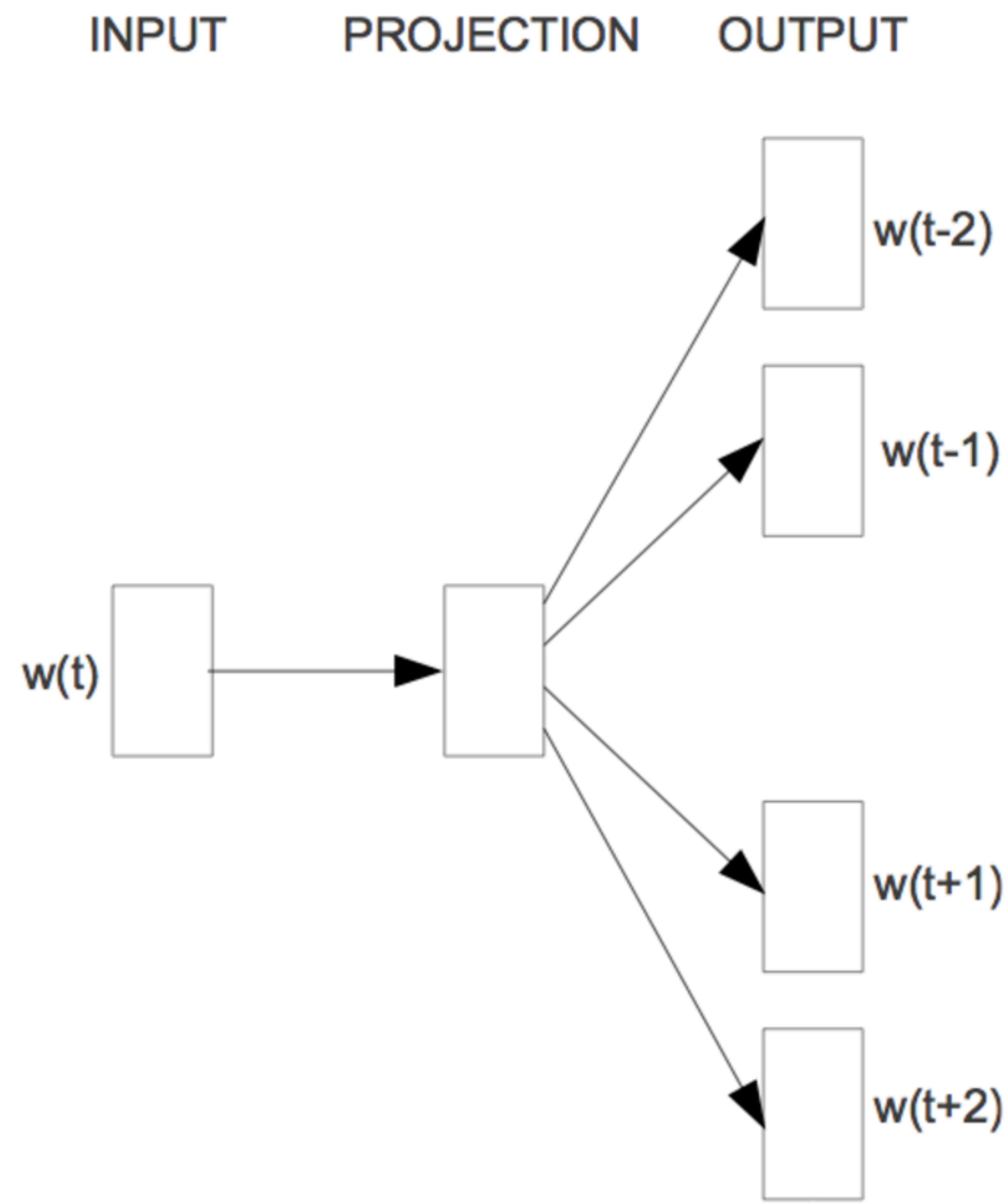
$$h = v_{w_t}$$

Objective to minimize:

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

Minimize this objective function with respect to

$$\theta = (v_w, v'_w \;\; \forall w)$$



Final vector representation
of word w : vec(w)

Option 1 $$vec(w) := v_w + v_w^{'}$$

Option 2 $$vec(w) := [v_w, v_w^{'}]$$