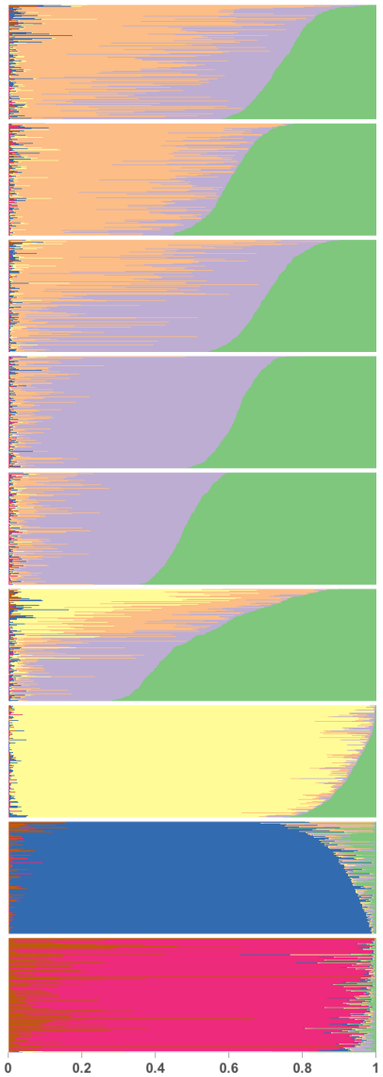
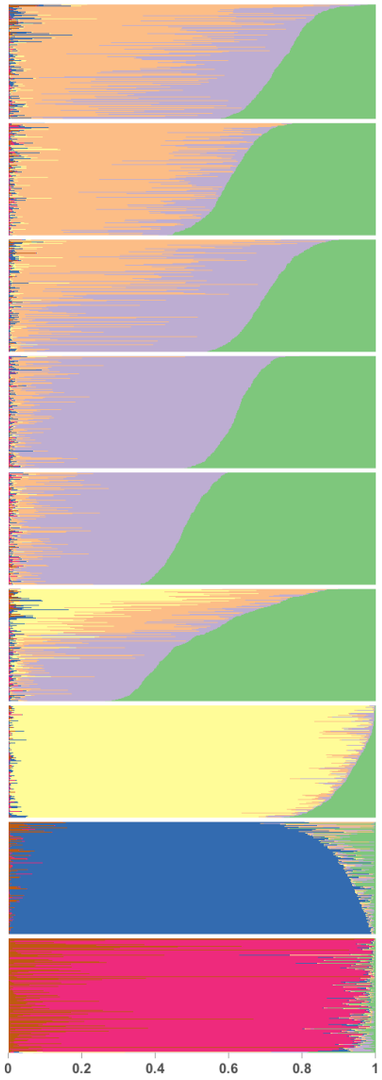
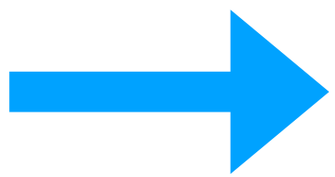
Lab Meeting

March 8, 2018

From STRUCTURE to *ash*







CorShrink

An adaptive method for correlation shrinkage

https://github.com/kkdey/CorShrink

gene expression data across 8555 tissue samples from 544 donors, across 51 tissues and 2 cell lines.

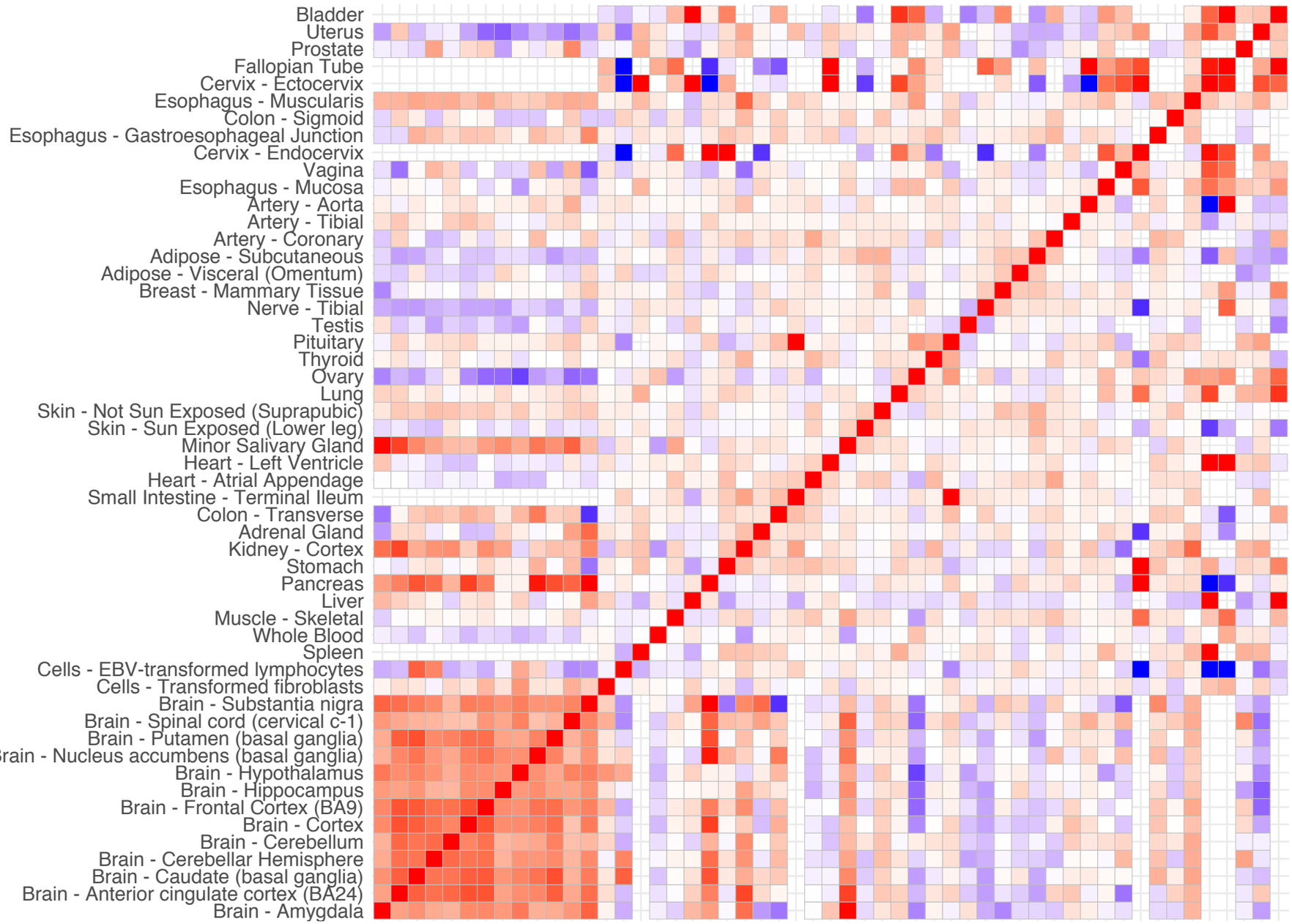
Example

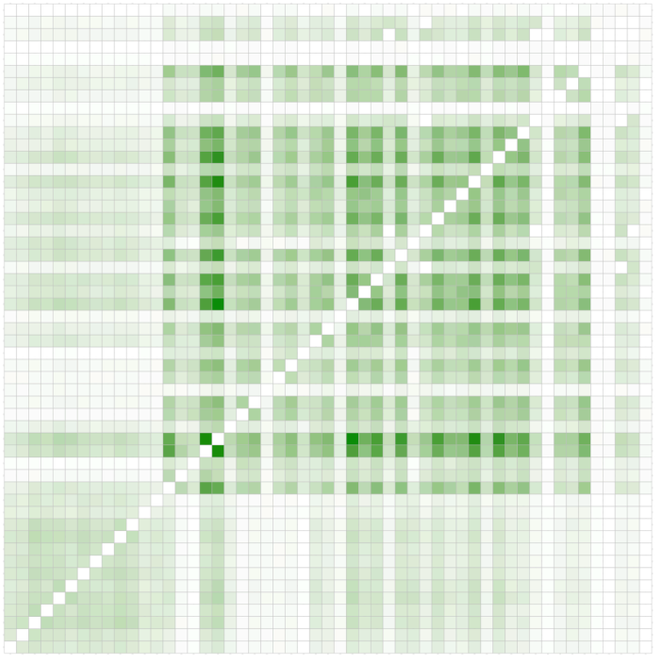
log CPM normalized expression for *PLIN1* gene across donors and tissues

```
> donor_by_tissue_PLIN1[1:5,1:5]
           Adipose - Subcutaneous Adipose - Visceral (Omentum) Adrenal Gland Artery - Aorta Artery - Coronary
GTEX-111CU              10.472332                     10.84006      2.721234             NA                 NA
GTEX-111FC               7.335392                           NA            NA             NA                 NA
GTEX-111VG               9.118889                           NA            NA             NA                 NA
GTEX-111YS              10.806459                     11.26113      3.454823       1.162059                 NA
GTEX-11220              11.040446                     11.71497      1.522667       1.674467           4.188002
```

Total proportion of missing observations : 70.3 %

Pairwise tissue-tissue correlation matrix (PLIN1 gene)

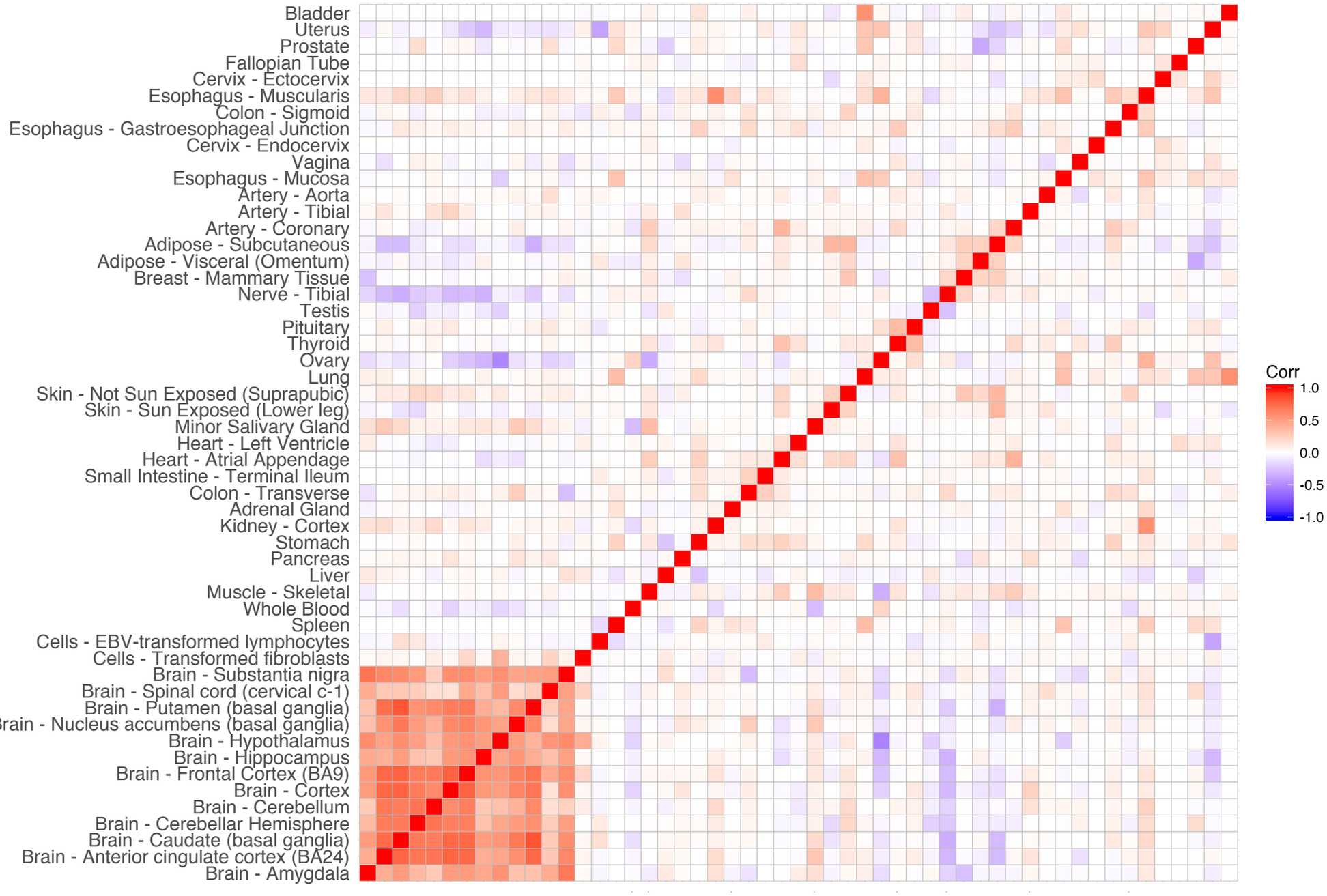




matched samples

Can we get a cleaner picture using a Correlation estimation method?

After *CorShrink* ……



Modeling Workflow

(R_{ij}, n_{ij})

R_{ij} : pairwise correlation tissues i and j

n_{ij} : number of matched donors between tissues i and j

$Z_{ij} = \frac{1}{2} log\left(\frac{1+R_{ij}}{1-R_{ij}}\right)$ $\quad$ $s_{ij} = \sqrt{\frac{1}{n_{ij}-1} + \frac{2}{(n_{ij}-1)^2}}$ $\quad$ biv. normal assumption

(Z_{ij}, s_{ij})

ash

$Z^{\star}_{ij}$

shrunk Fisher z-score

$R^{\star}_{ij} = \frac{exp(2Z^{\star}_{ij}-1)}{exp(2Z^{\star}_{ij}+1)}$

$R^{\star}_{ij}$

shrunk correlations

The matrix $R^{\star}$ may not be PD, so we opt for the nearest PD matrix $R^{\star\star}$ (Higham 2002)

What does ash do?

$$Z_{ij} \sim N(\eta_{ij}, s_{ij})$$

η_{ij} : population Fisher z-score between tissues i and j

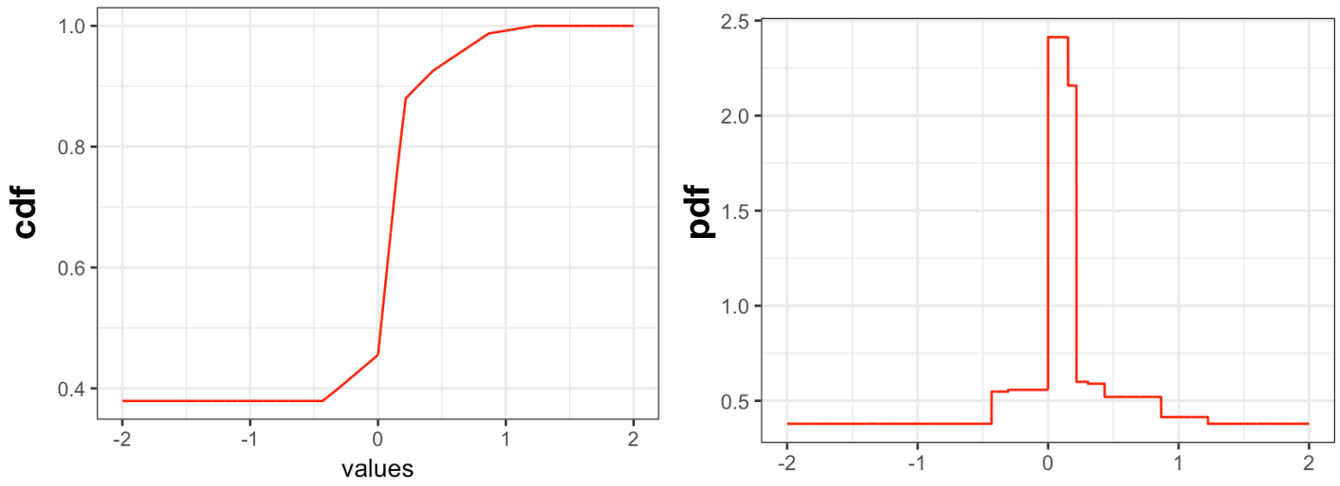
$$\eta_{ij} \sim g(\eta) = \sum_{k=1}^{K} \pi_k N\left(0, \sigma_k^2\right) \qquad normal$$

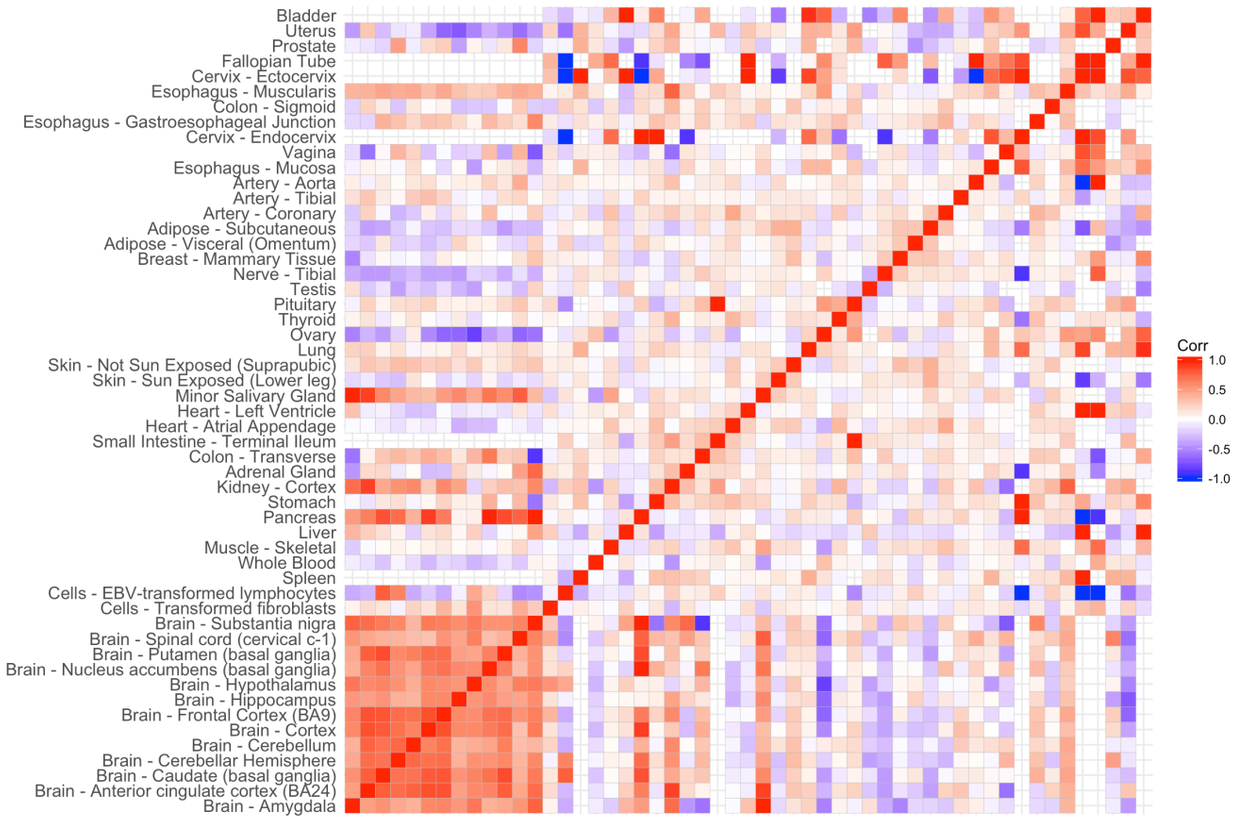
$$\eta_{ij} \sim g(\eta) = \sum_{k=1}^{K} \pi_k \{U(-a_k, 0) \mid U(0, b_k)\} \qquad halfuniform$$

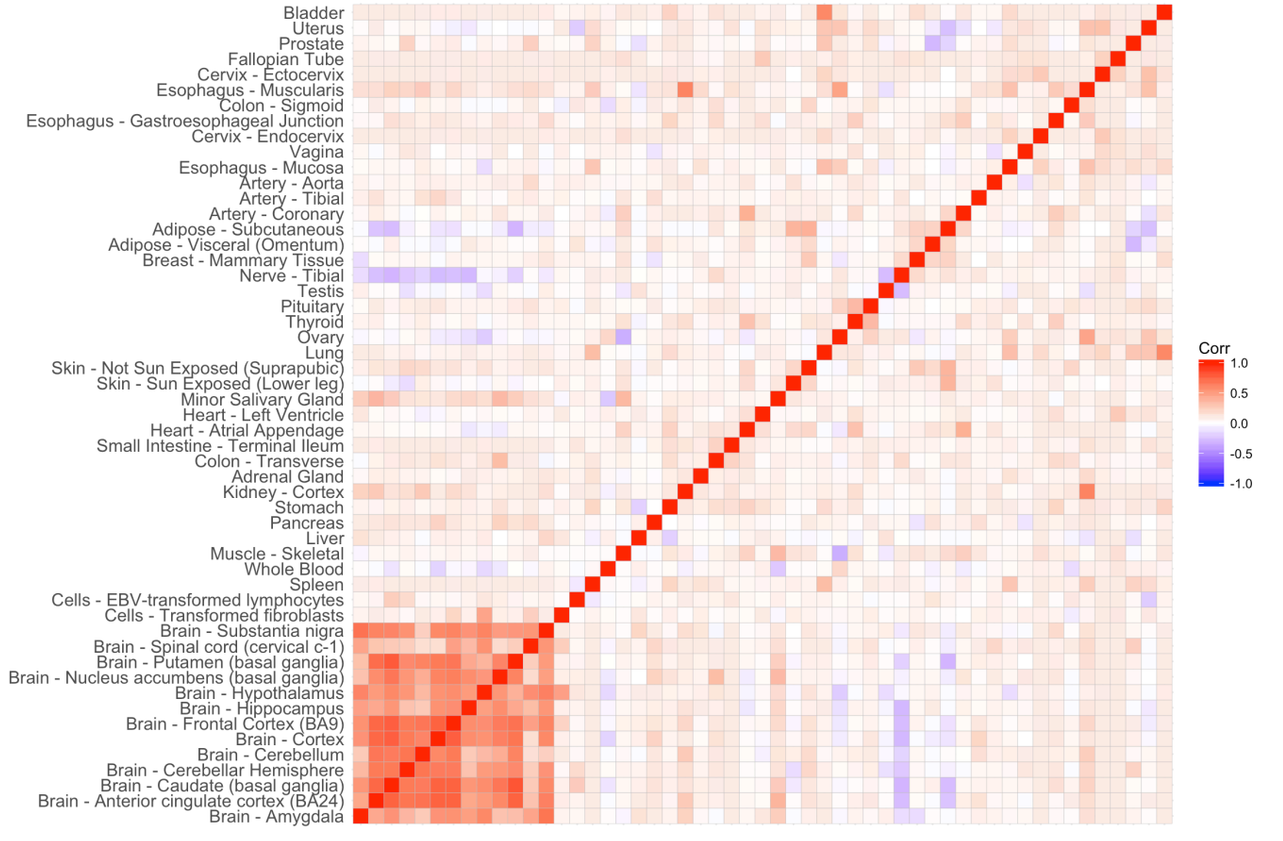
The mixture prior is unimodal but its model may be different from 0.

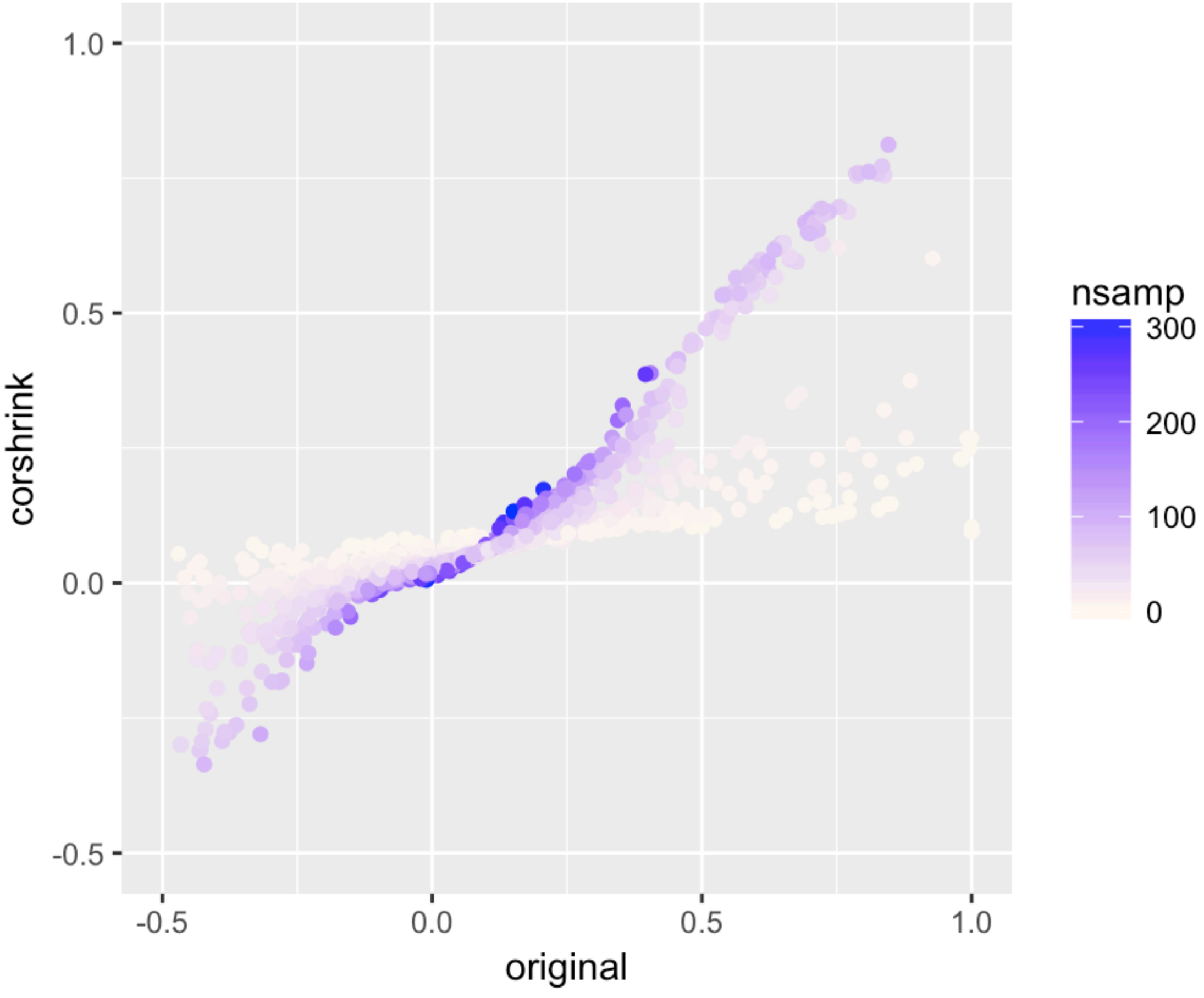
$$Z_{ij}^{\star} = E\left(\eta_{ij} | Z_{ij}, s_{ij}\right) = E\left(\eta_{ij} | R_{ij}, n_{ij}\right)$$

$$g(\eta)$$


cdf
1.0
0.8
0.6
0.4
-2
-1
0
1
2
values
pdf
2.5
2.0
1.5
1.0
0.5
-2
-1
0
1
2


Bladder
Uterus
Prostate
Fallopian Tube
Cervix - Ectocervix
Esophagus - Muscularis
Colon - Sigmoid
Esophagus - Gastroesophageal Junction
Cervix - Endocervix
Vagina
Esophagus - Mucosa
Artery - Aorta
Artery - Tibial
Artery - Coronary
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Breast - Mammary Tissue
Nerve - Tibial
Testis
Pituitary
Thyroid
Ovary
Lung
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Minor Salivary Gland
Heart - Left Ventricle
Heart - Atrial Appendage
Small Intestine - Terminal Ileum
Colon - Transverse
Adrenal Gland
Kidney - Cortex
Stomach
Pancreas
Liver
Muscle - Skeletal
Whole Blood
Spleen
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Brain - Substantia nigra
Brain - Spinal cord (cervical c-1)
Brain - Putamen (basal ganglia)
Brain - Nucleus accumbens (basal ganglia)
Brain - Hypothalamus
Brain - Hippocampus
Brain - Frontal Cortex (BA9)
Brain - Cortex
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Brain - Anterior cingulate cortex (BA24)
Brain - Amygdala
Corr
1.0
0.5
0.0
-0.5
-1.0


Bladder
Uterus
Prostate
Fallopian Tube
Cervix - Ectocervix
Esophagus - Muscularis
Colon - Sigmoid
Esophagus - Gastroesophageal Junction
Cervix - Endocervix
Vagina
Esophagus - Mucosa
Artery - Aorta
Artery - Tibial
Artery - Coronary
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Breast - Mammary Tissue
Nerve - Tibial
Testis
Pituitary
Thyroid
Ovary
Lung
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Minor Salivary Gland
Heart - Left Ventricle
Heart - Atrial Appendage
Small Intestine - Terminal Ileum
Colon - Transverse
Adrenal Gland
Kidney - Cortex
Stomach
Pancreas
Liver
Muscle - Skeletal
Whole Blood
Spleen
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Brain - Substantia nigra
Brain - Spinal cord (cervical c-1)
Brain - Putamen (basal ganglia)
Brain - Nucleus accumbens (basal ganglia)
Brain - Hypothalamus
Brain - Hippocampus
Brain - Frontal Cortex (BA9)
Brain - Cortex
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Brain - Anterior cingulate cortex (BA24)
Brain - Amygdala
Corr
1.0
0.5
0.0
-0.5
-1.0

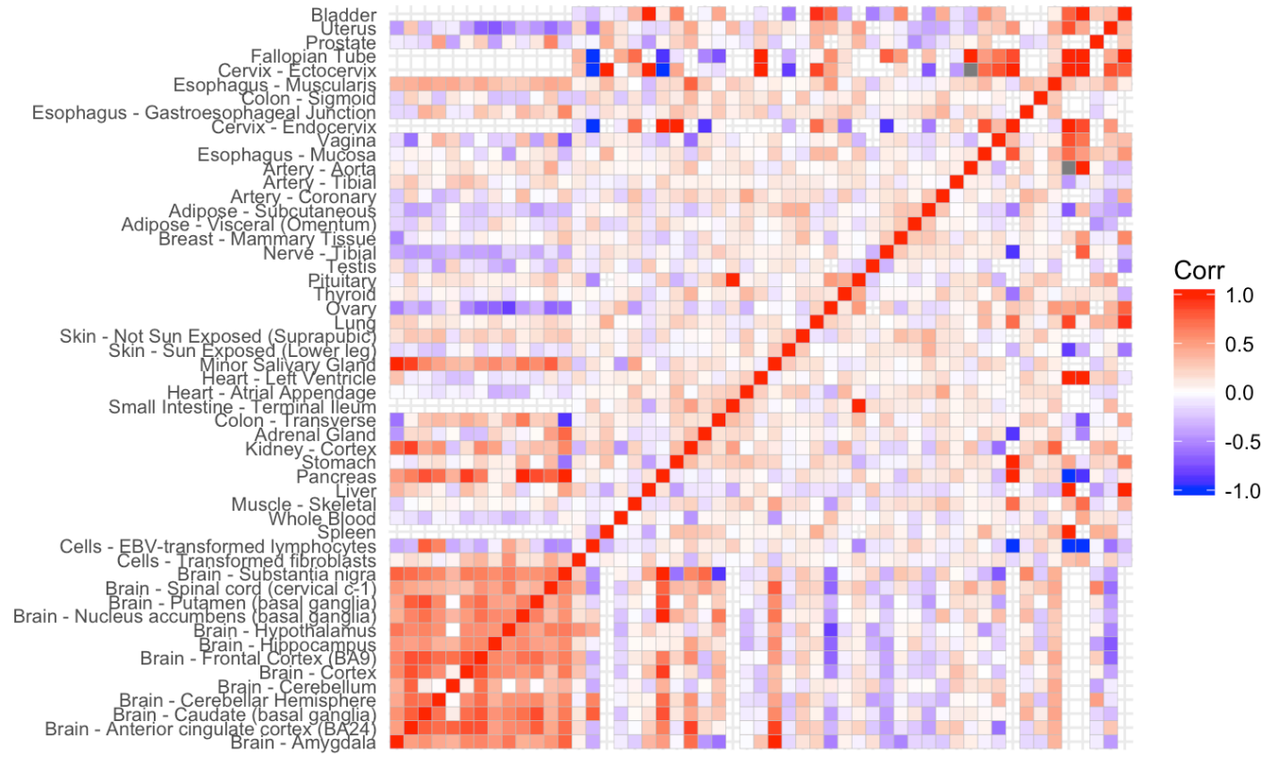

1.0
0.5
0.0
-0.5
corshrink
-0.5
0.0
0.5
1.0
original
nsamp
300
200
100
0

CorShrink

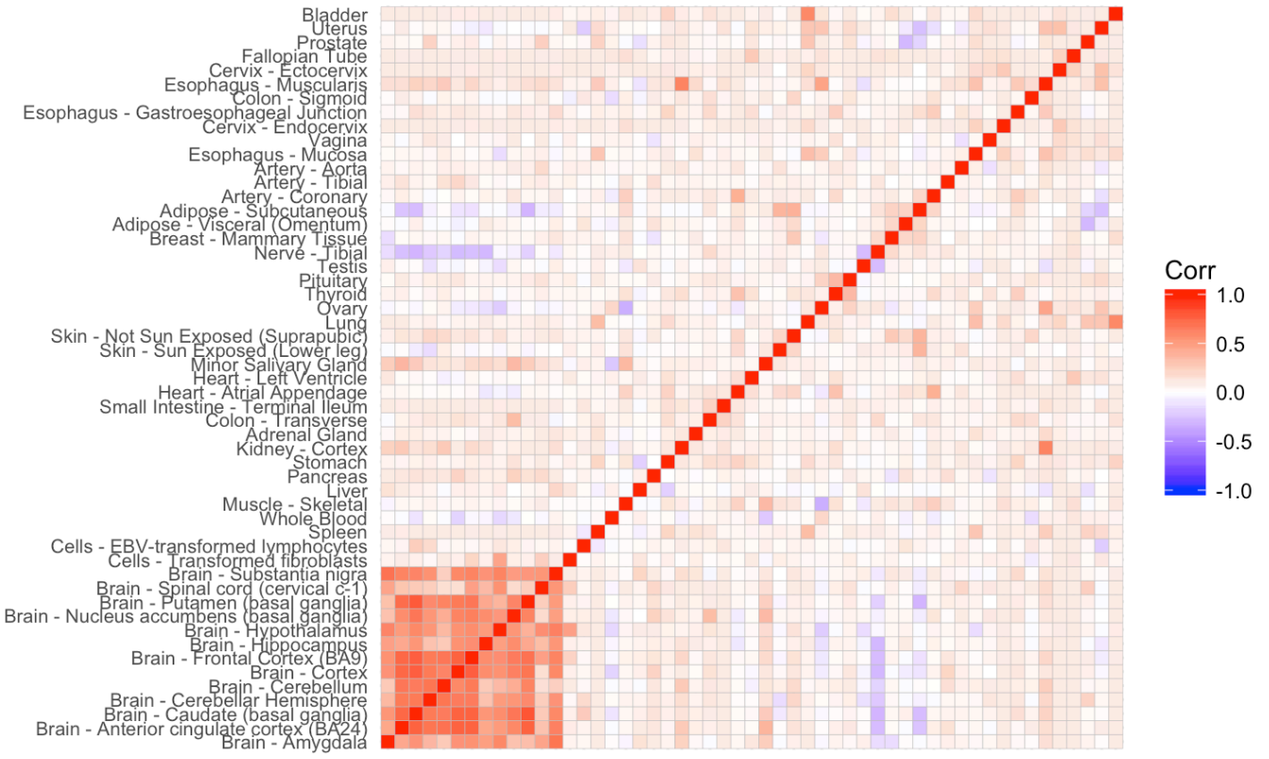
vs

Imputation

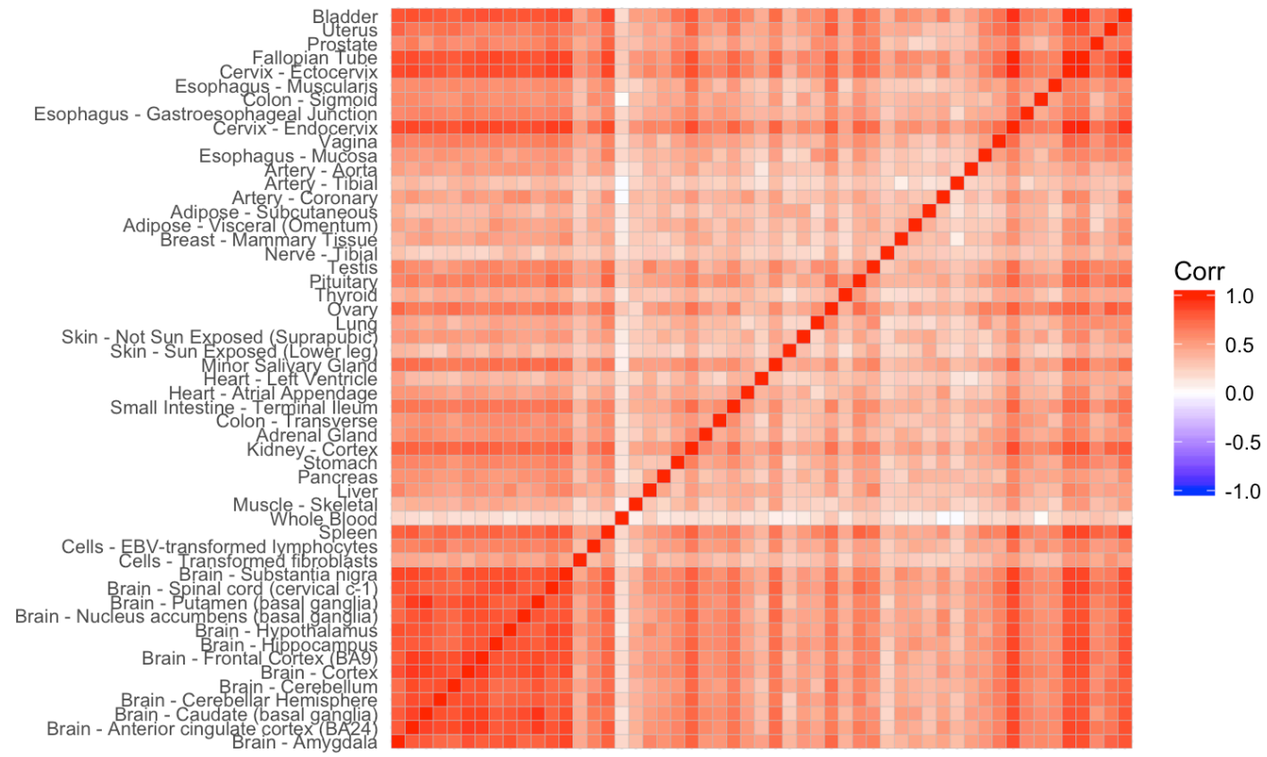
pairwise correlation



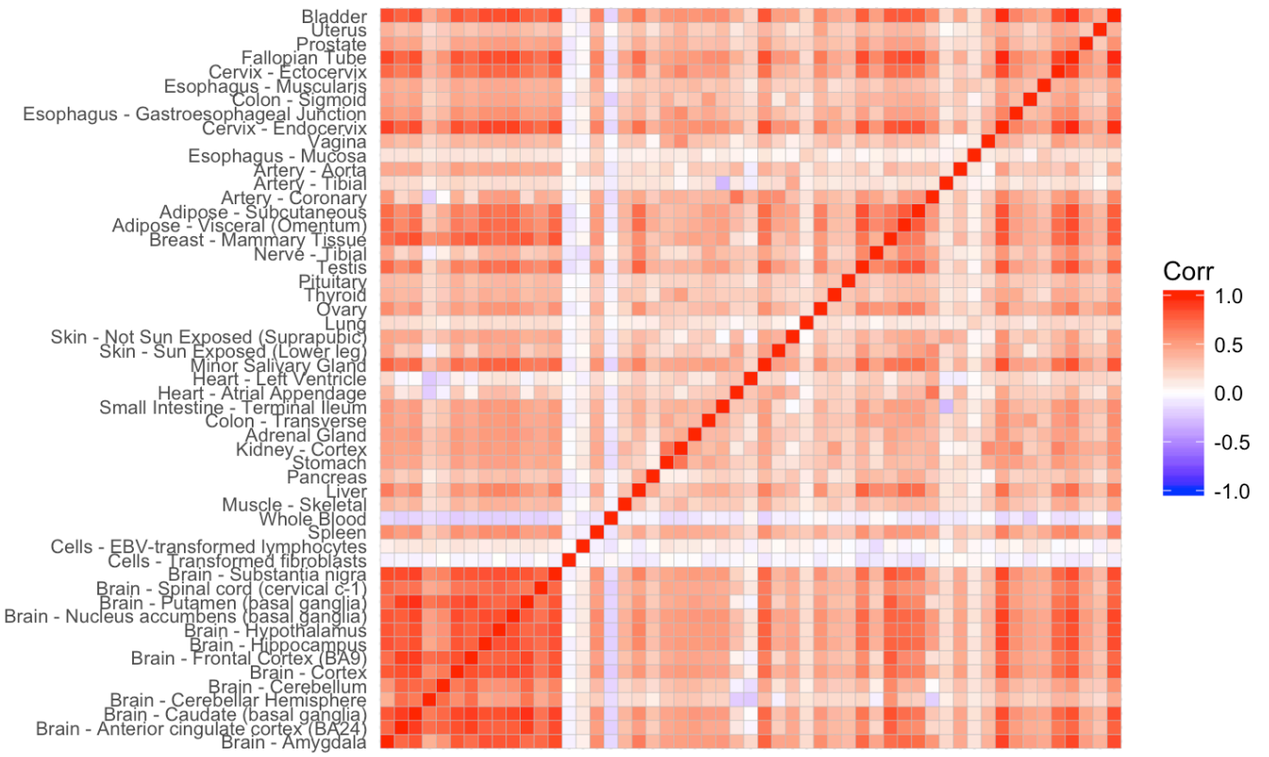
CorShrink



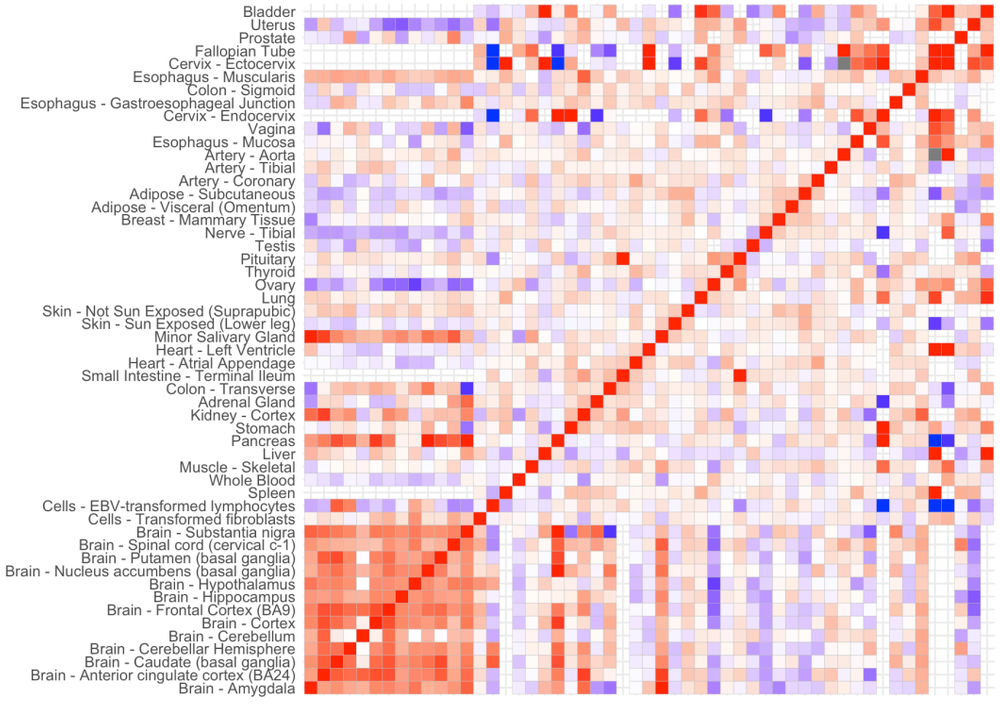
correlation after SoftImpute imputed data



correlation after FLASH imputed data

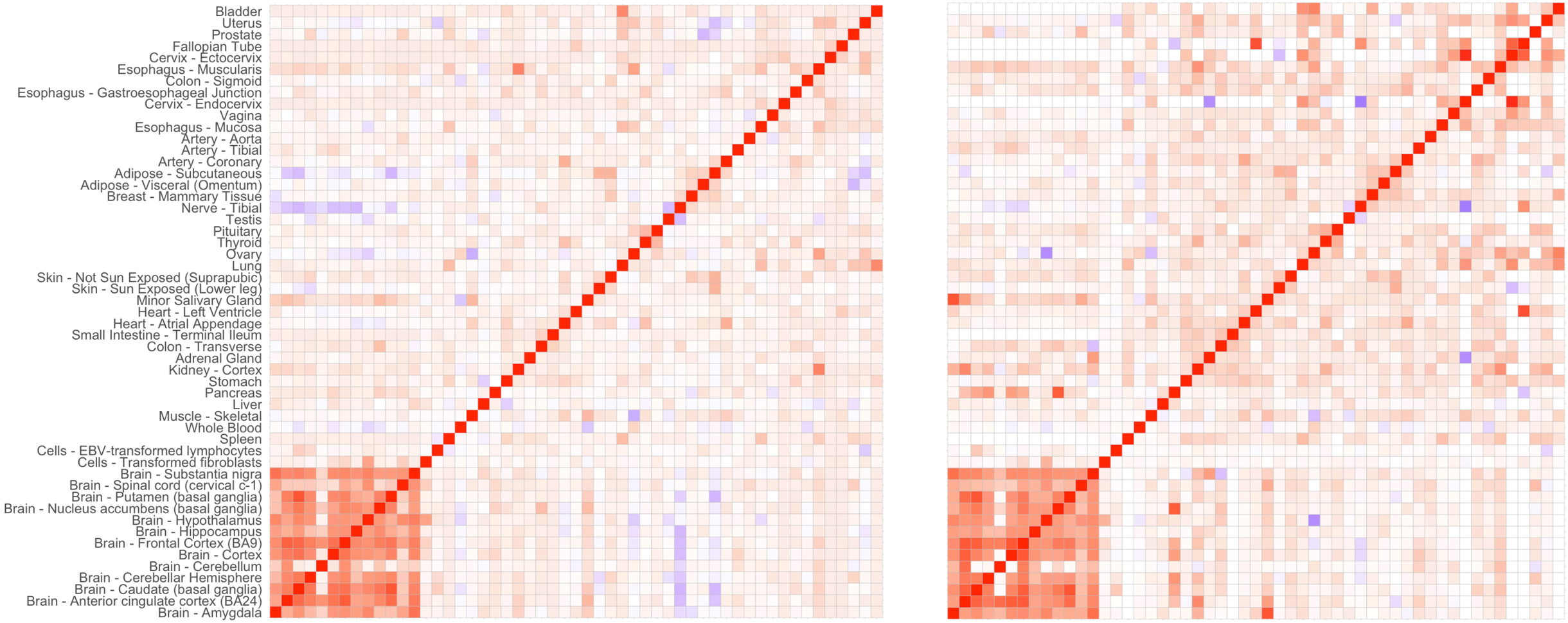


Tissue-wide
vs
Gene-wide
CorShrink

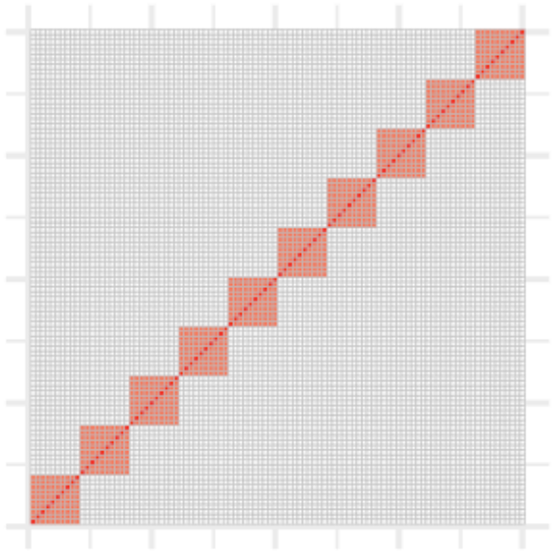
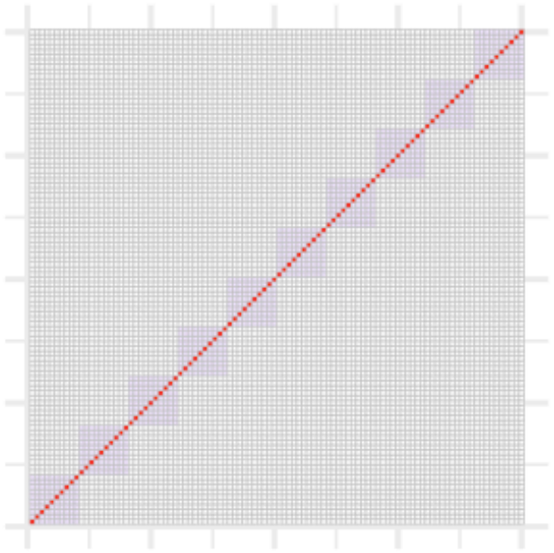
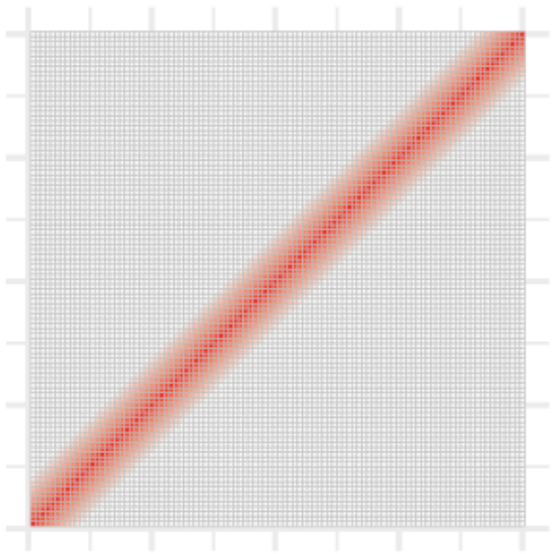
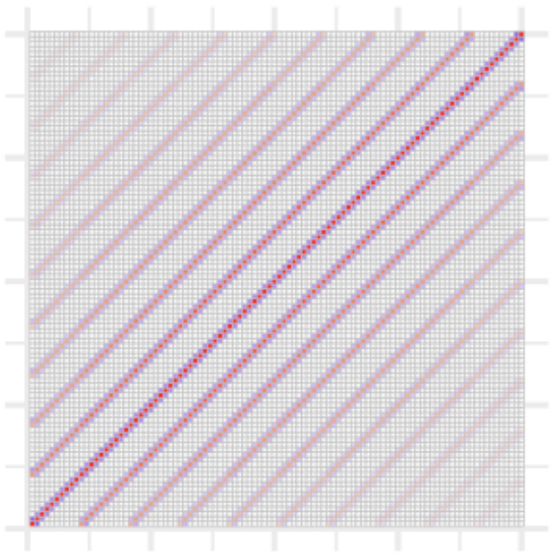
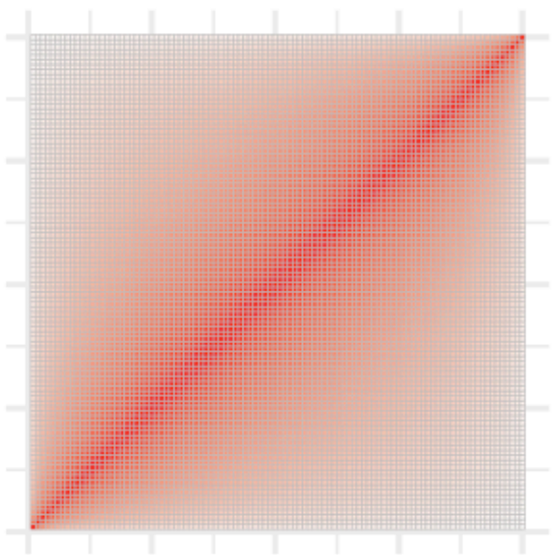
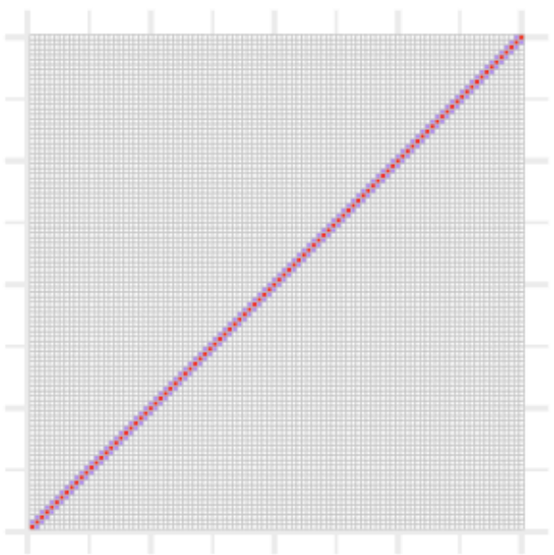


tissue wide CorShrink

gene wide CorShrink



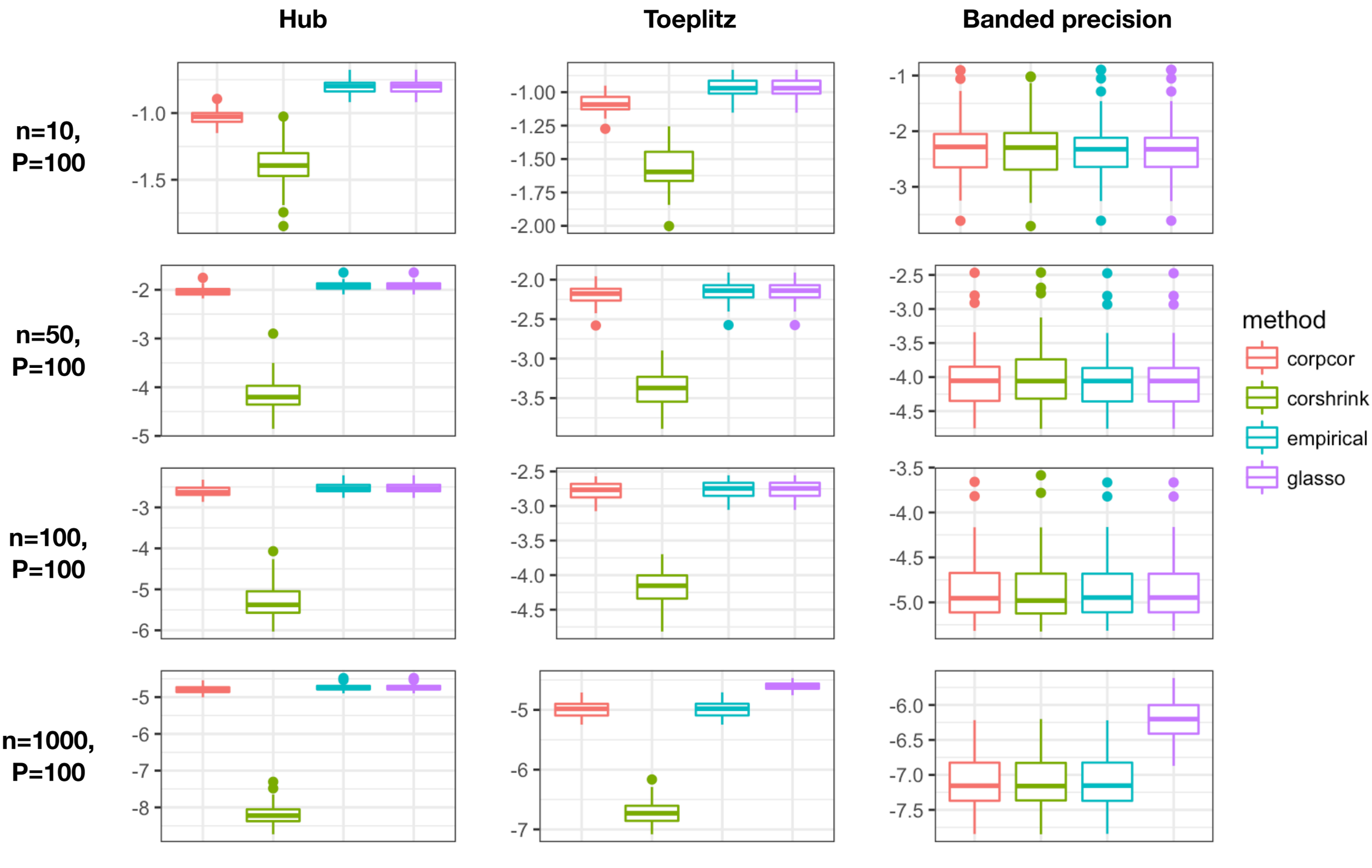
Simulation Studies

	correlation	inverse correlation
Hub correlation		
Toeplitz correlation		
Banded precision		

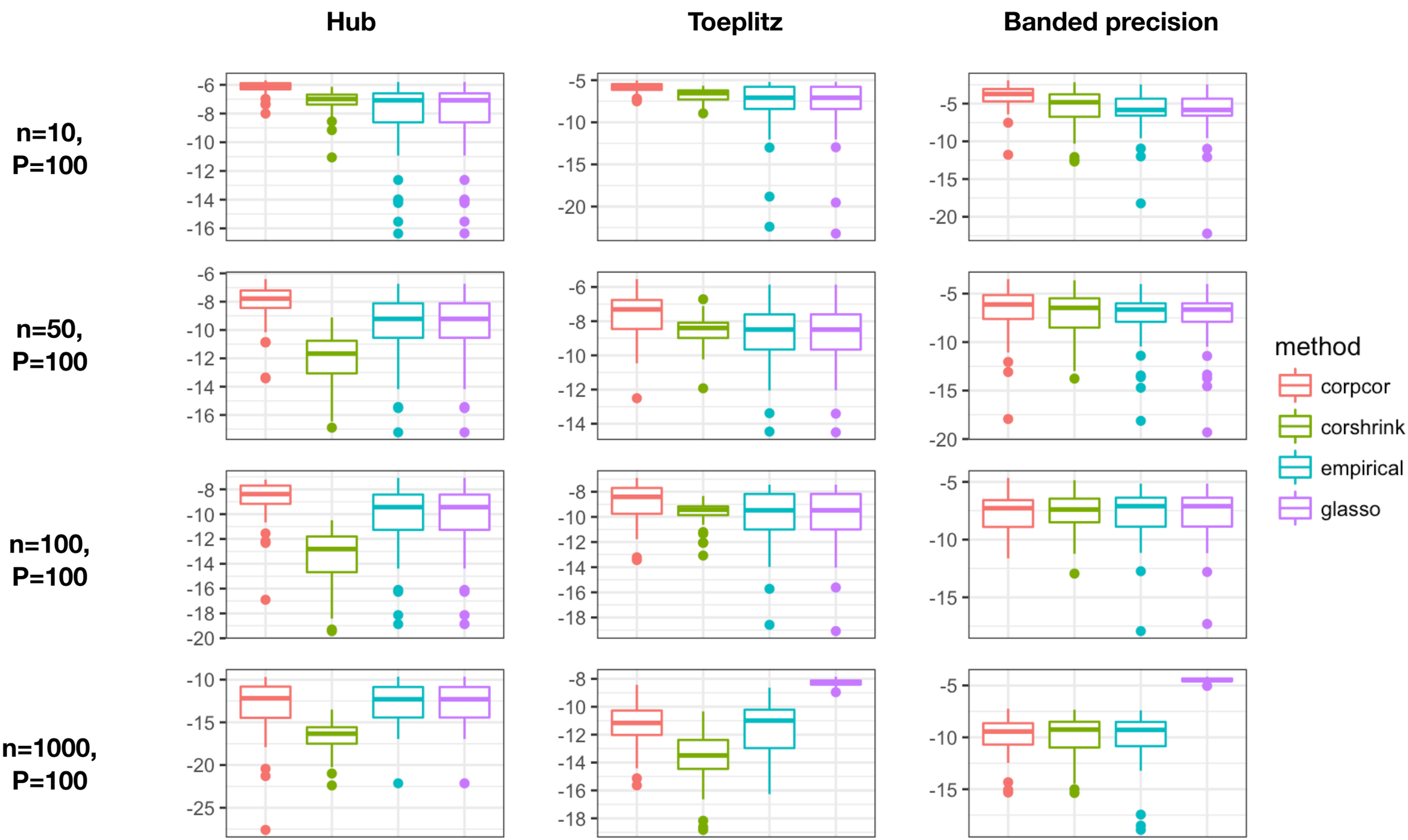
color grade

Correlation matrix distance (CMD) :

1 - (cosine similarity between vectorized correlations between two correlation matrices)



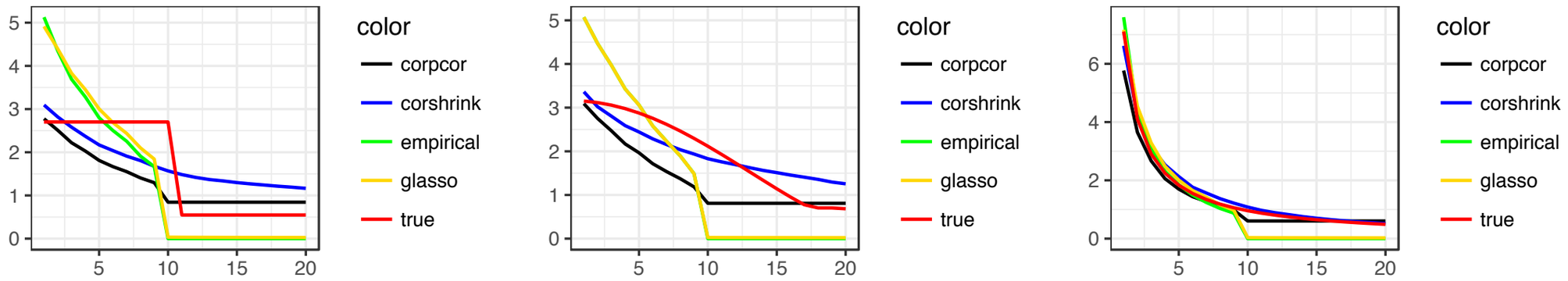
Frobenius Distance



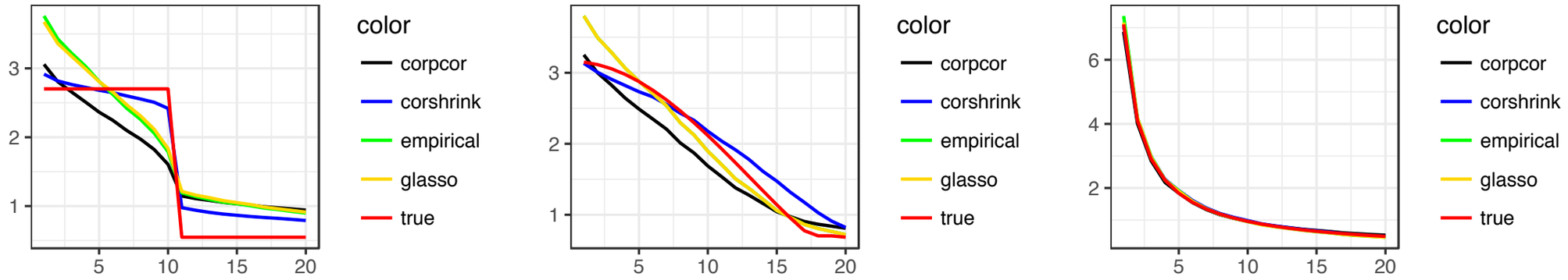
Eigenvalue Distributions (square root) scale

Hub | **Toeplitz** | **Banded precision**

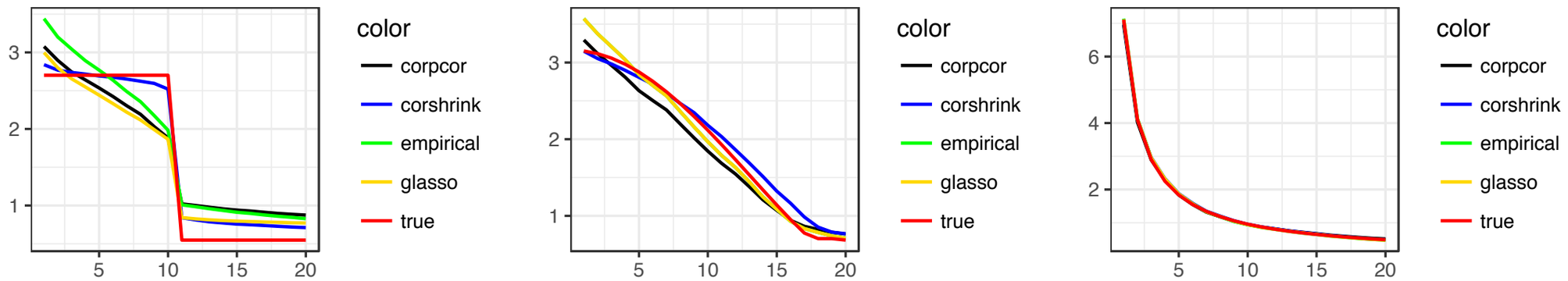
n=10, P=100



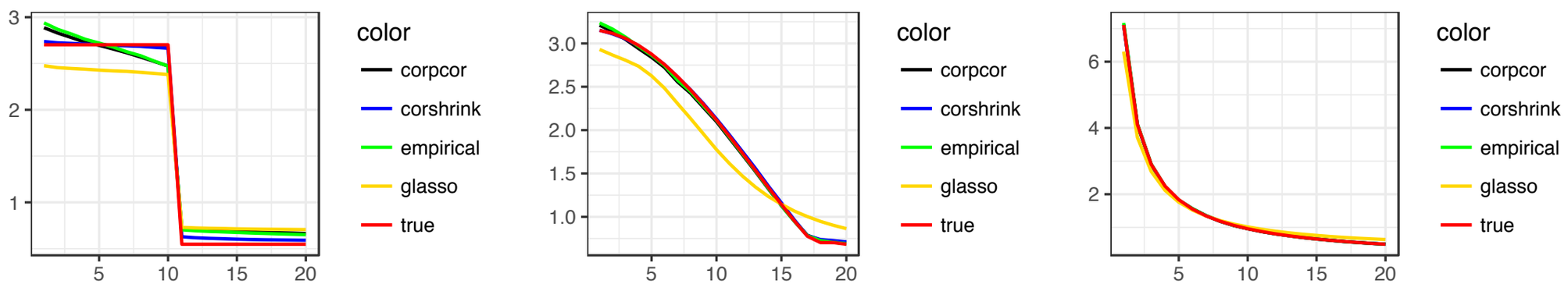
n=50, P=100



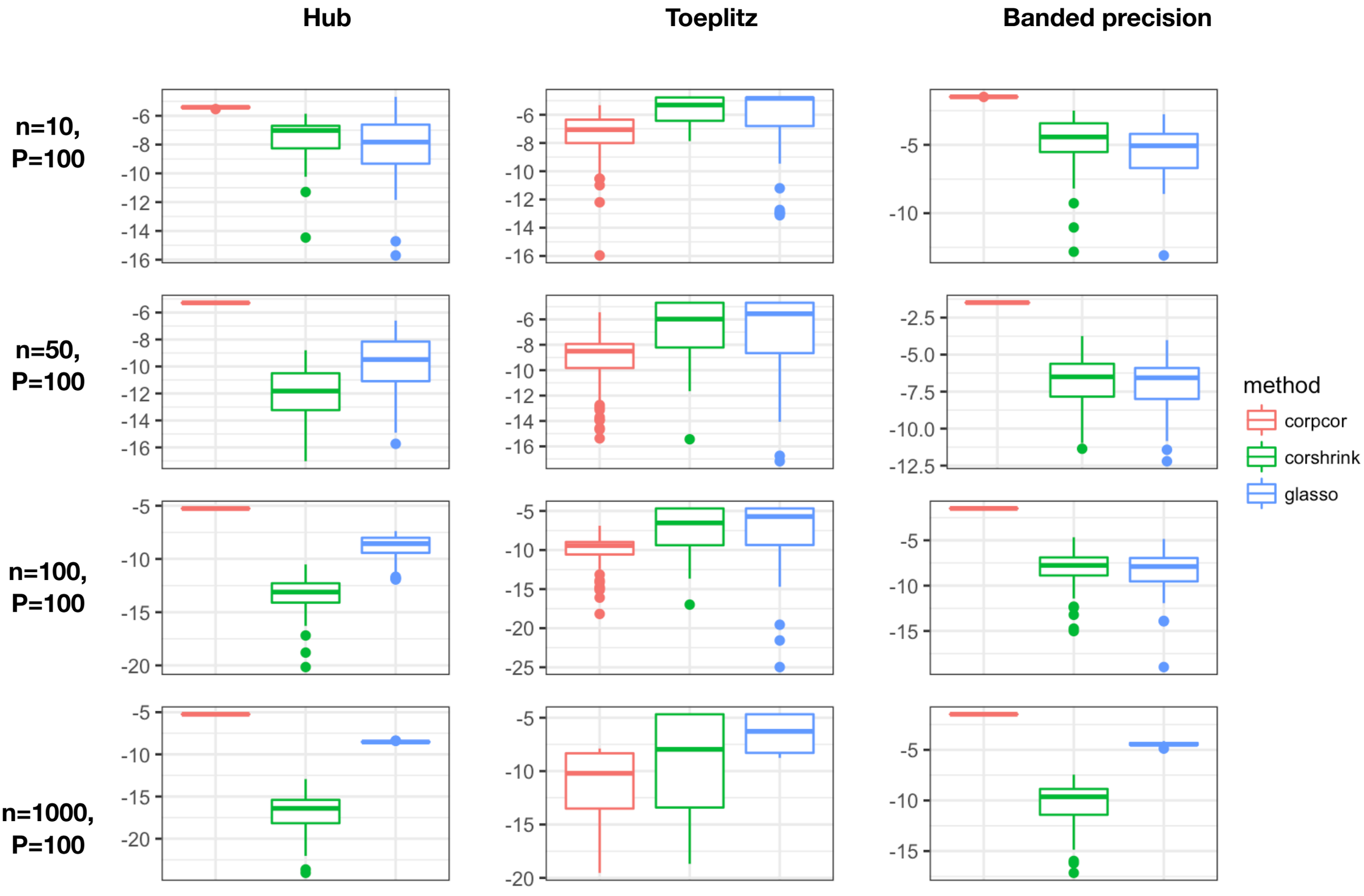
n=100, P=100



n=1000, P=100



Inverse Correlation Comparison (Frobenius norm)



CorShrink on word similarities

1968 America

civil rights movement


GOOD MORNING
Although the area skies will clear today, the weather will be clouded by a chilly spell.

MORNING ADVOCATE

THE INDEX
Amusements 8-D Financial ...15-3
Classified ... 9E Society 3-B
Comics .... 2E Sports 2-C
Deaths .... 12-A Television ..10-B

43rd Year, No. 279 ★ Baton Rouge, La., Friday Morning, April 5, 1968 Associated Press, United Press International 66 Pages Ten Cents

King Shot to Death in Memphis

Relief Forces Push Close to Khe Sanh

SAIGON (AP)—A U.S. task force pressed close to the Marine combat base at Khe Sanh on Friday, the fifth day of a drive to lift the long siege of the battered fortress in South Vietnam's northwest corner.

At last report, lead elements of the relief force were within half a mile Thursday night of the Khe Sanh perimeter and were poised for a final push.

The relief force has met only light opposition from artillery and mortar fire, mainly jumping off from Co No. 12 miles away. On Monday, Soviet reporter in Hanoi said Thursday that the North Vietnamese were beginning to withdraw as a good-will gesture, but there was no confirmation from Hanoi.

Another reason advanced for the lack of sizeable opposition was the incessant American bombing by planes and from artillery in Khe Sanh itself, as North Vietnamese positions during the past three months.

In other developments early Friday:

Allied military spokesmen said the Viet Cong shelled 37 U.S. and South Vietnamese military positions Thursday night and early Friday. The targets, ranging the length of the country, included the big Tan Son Nhut air base on Saigon's outskirts. The heaviest shelling was at a U.S. Army base camp...

Reports Say McKeithen May Ask Fee

Gov. McKeithen may ask the legislature to raise the automobile license tag fee without proposing a constitutional amendment, the attorney general's office said Thursday.

First Asst. Atty. Gen. Carroll Bush said in an opinion requested by the governor that the Legislature may legally impose a license tax in excess of the $5 now charged annually.

The opinion added that all additional revenue from the tax must go into the long-range highway fund.

The ruling cleared the way for McKeithen to ask lawmakers for a plan to finance a massive new road building program planned by the department.

Highway officials are calling for $70 million in new funds annually to finance their program...

DR. KING SLAIN—Dr. Martin Luther King was killed Thursday in Memphis, Tenn., in a downtown hotel. He was taken to a Memphis hospital after being shot in the neck and was pronounced dead. —AP wirephoto

Curfew Reimposed; Guard Sent Back

MEMPHIS, Tenn. (AP)—Nobel Laureate Martin Luther King Jr., father of non-violence in the American Civil Rights movement, was killed by an assassin's bullet Thursday night.

King, 39, was hit in the neck by a bullet as he stood on the balcony of a motel here. He died less than an hour later in St. Joseph Hospital.

Gov. Buford Ellington immediately ordered 4,000 National Guard troops back into the city. A curfew, which was clamped on Memphis after a King-led march turned into a riot a week ago, was reimposed.

* * *

Police said incidents of violence, including several fire bombings were reported following King's death.

The 1964 Nobel Peace Prize winner, was standing on the balcony of his motel here, where he had come to lead protests in behalf of the city's 1,300 striking garbage workers, most of them Negroes, when he was shot.

Two unidentified men who were arrested were released several hours later.

As word of King's death spread through the stunned city, Negroes in scattered areas stoned stores, stoned police and fire trucks and tossed several fire bombs. Two policemen were injured, mainly by rioters...

Louisiana Negroes React

By BARRY ZANDER

The death of Dr. Martin Luther King will "definitely put a strain on Louisiana race relations" was the opinion Thursday night expressed by Emmitt Douglas, leader of the state chapter of the National Association for the Advancement of Colored People, and echoed by many civil rights move-

LBJ Postpones Trip to Hawaii

WASHINGTON (AP)—President Johnson, preparing for a weekend Vietnam strategy conference in Honolulu, delayed his departure overnight Thursday because of the assassination in Memphis of the Rev. Dr. Martin Luther King.

The President went before television cameras and radio microphones to declare America is "shocked and saddened" by the tragedy. He condemned violence, lawlessness and divisiveness.

Johnson goes to Honolulu amid growing confusion here about the prospects of preliminary peace talks.

The White House did not indicate whether Johnson would change his earlier plan to meet at March Air Force Base, Calif., to talk with former President Dwight D. Eisenhower.

The original plan called for Johnson also to meet in Hawaii Sunday with South Korea's president, Chung Hee Park.

Earlier in the day, Hanoi charged that American planes bombed a North Vietnamese city...

(Continued on Page 10-A, Col. 1)

Governor, HHH Hold Confidential Talk

NEW ORLEANS (AP)—Gov. McKeithen revealed Thursday he had a private, half-hour telephone conversation with Vice President Hubert H. Humphrey, but their discussion was "confidential."

The governor would not say if a possible Humphrey - McKeithen ticket for president was in the making. However, he said, "I certainly would consider being a vice presidential candidate with Humphrey."

Satellite Is Hurled Into Wrong Orbit

assassination of Dr. Martin Luther King Jr.

vietnam war

anti war protests

DNC protests


City Edition

Eugene Register-Guard

Some Sun Friday

101st Year, No. 227 FOUR SECTIONS Eugene, Oregon, Thursday, June 6, 1968 FIFTY-SIX PAGES Price 10 Cents

Kennedy Dies From Wounds

Flags Lowered

LBJ Sets Day Of Mourning

Senator to Be Buried In Arlington Cemetery

Death Brings Moratorium To Campaign

House Votes Gun Sale Restrictions

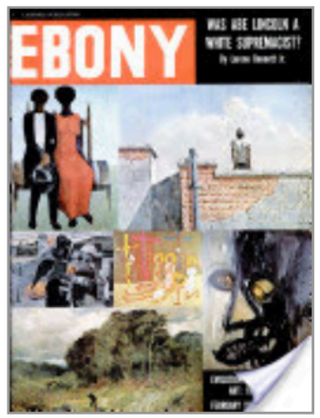
RFK's View Of Dangers Recalled

RFK assassination

Text data from Ebony magazines for 1968

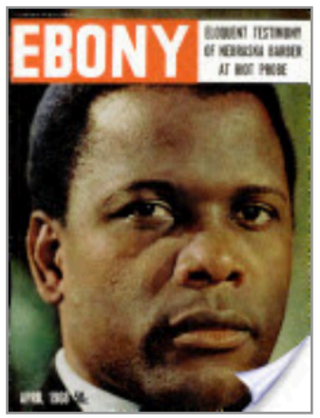
(magazines available under Google books)

Jan 1968

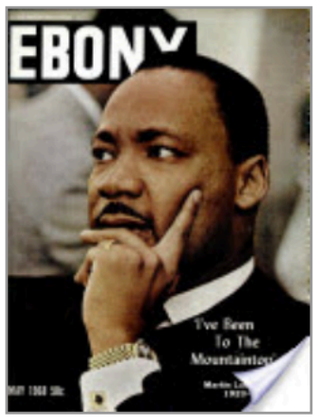


Feb 1968

Mar 1968



Apr 1968



May 1968

Jun 1968

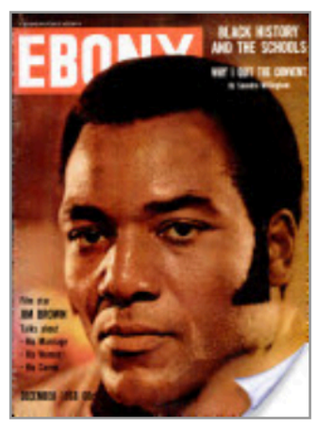
Jul 1968

Aug 1968

Sep 1968

Oct 1968

Nov 1968

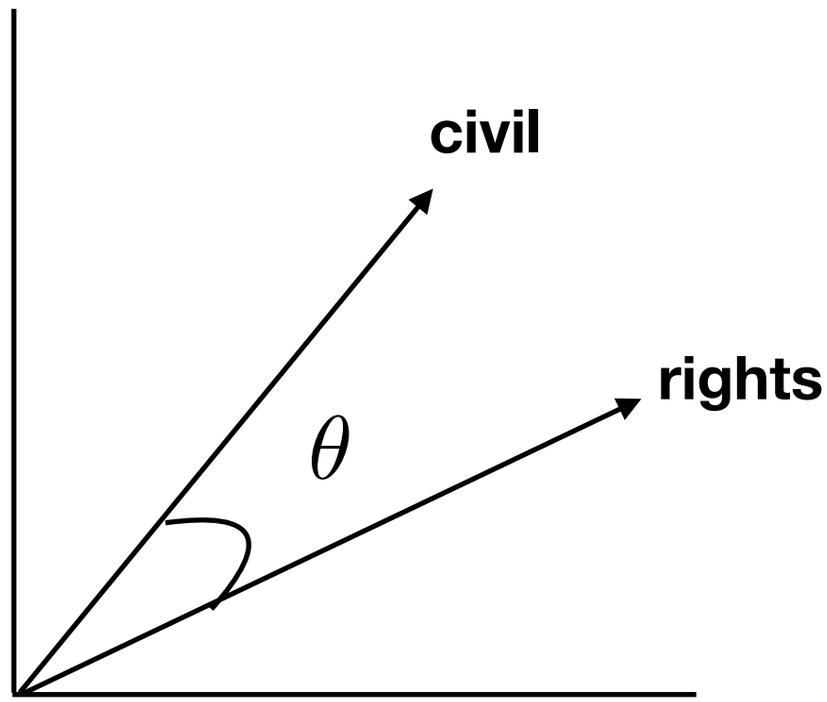


Dec 1968

I fitted *word2vec* model on the combined text data for all monthly issues.

Word2vec Model

A neural network model that uses the neighbors of each word to learn a vector representation of the word.



Cosine similarity between the words - *civil* and *rights* given by the cosine of the angle between their vector representations

$$cos(\theta)$$

CorShrink + word2vec

Fix a word or a set of words we are interested in (say civil, rights)

Consider N words closest to the word(s) of interest as per *word2vec* model.

Consider the pairwise cosine similarities of each of the $N(N-1)/2$ pairs

Convert these cosine similarities to Fisher z-score type measure

CorShrink + word2vec

Fix a word or a set of words we are interested in (say civil, rights)

Consider N words closest to the word(s) of interest as per *word2vec* model.

Consider the pairwise cosine similarities of each of the $N(N-1)/2$ pairs

Convert these cosine similarities to Fisher z-score type measure

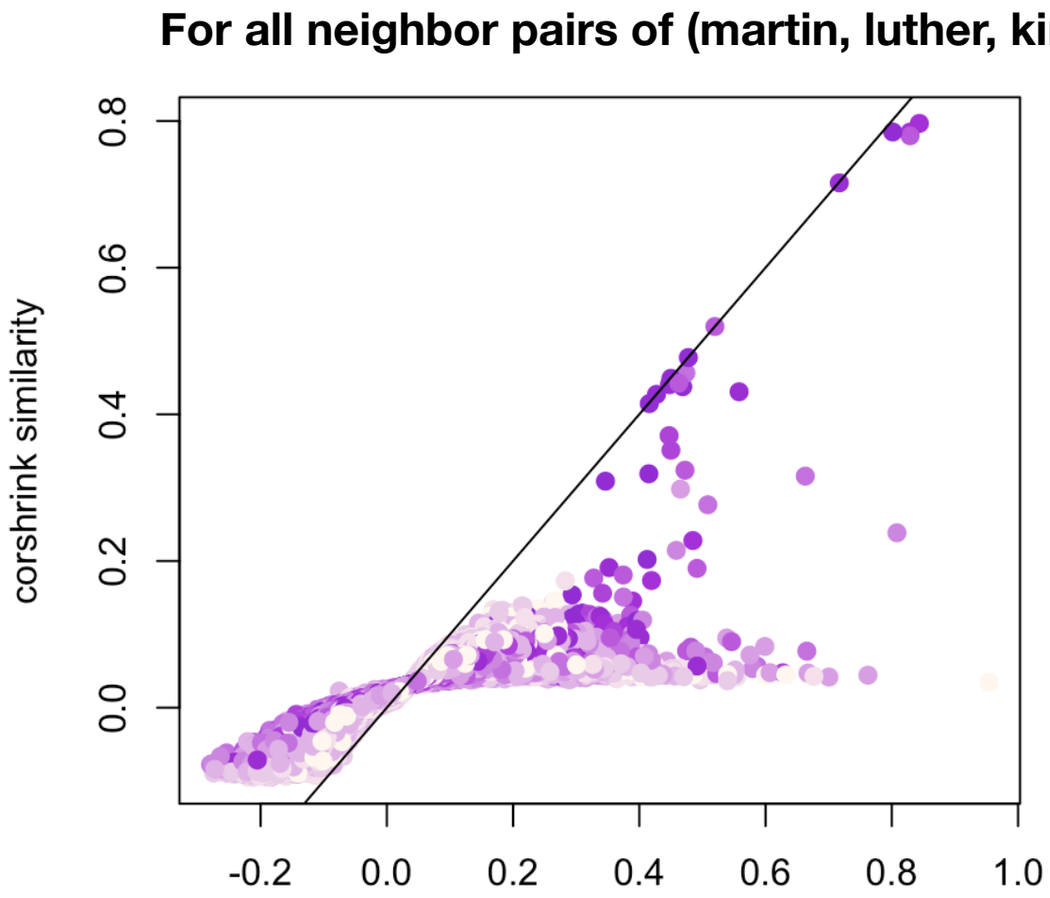
How to calculate the standard errors of these z-scores?

Resample the magazine issues 100 times in parallel and fit *word2vec*, obtain cosine similarities and compute Fisher z-scores across pairs for each re-sample

Compute the standard error of the resampled Fisher z-scores between a pair of words across 100 re-samples

No modeling assumption required; but subject to Bootstrap error

Words close in context to Dr. Martin Luther King Jr.



before *CorShrink*

(nearest 25 words to martin, luther, king)

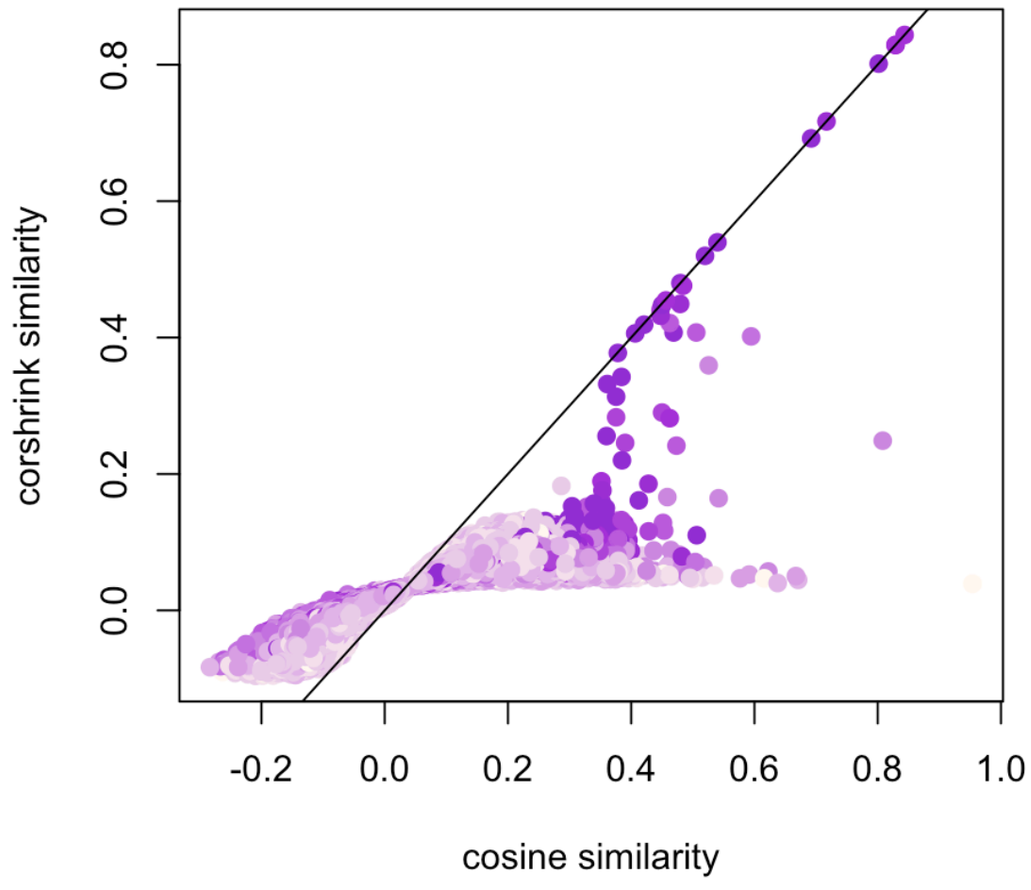
luther	martin	king	rev	apostle	floyd	requested
0.8908471	0.8535049	0.8485832	0.4063000	0.3981613	0.3509520	0.3498945
fiery	assassinated	murder	francis	funeral	forres	late
0.3463924	0.3427114	0.3375902	0.3301637	0.3293240	0.3269475	0.3180746
kings	joan	prince	caucuses	abernathy	preaching	marched
0.3070072	0.3051331	0.3048364	0.2954050	0.2925116	0.2891073	0.2868515
prophet	kennedy	assassination	seeds			
0.2835925	0.2812839	0.2809260	0.2785261			

after *CorShrink*

(nearest 25 words to martin, luther, king)

luther	martin	king	rev	jackson	leader	assassination
0.86047946	0.83742984	0.83344445	0.23006744	0.10901378	0.09842106	0.09665314
james	kings	ralph	montgomery	courtship	late	birmingham
0.09482433	0.09374696	0.09264236	0.09258621	0.09234065	0.09025221	0.09022915
murder	civil	peace	actively	wright	personal	rights
0.08954815	0.08924835	0.08922241	0.08921495	0.08874579	0.08867831	0.08789465
naacp	god	voice	marched			
0.08683455	0.08661692	0.08588993	0.08563276			

Words close in context to civil and rights



before *CorShrink*

(nearest 25 words to civil and rights)

civil	rights	movement	legislation	disorders	bills
0.9007841	0.9007841	0.4405455	0.3730043	0.3550964	0.3421801
enforcement	movements	strengthened	involvement	equal	protection
0.3402390	0.3323286	0.3269478	0.3155524	0.3120888	0.3120654
reconstruction	belonged	cowboys	andrew	randolph	amendments
0.3117899	0.3022773	0.2964739	0.2961406	0.2911620	0.2899793
murders	commission	nonviolent	recommendations	brutal	sncc
0.2860052	0.2846442	0.2844914	0.2823338	0.2760409	0.2733980
paradox					
0.2722072					

after *CorShrink*

((nearest 25 words to civil and rights)

civil	rights	movement	equal	militant	commission	war
0.90078411	0.90078411	0.28434679	0.11712699	0.11265973	0.11157274	0.10768549
freedom	committee	involvement	luther	equality	principles	became
0.10677891	0.10644511	0.10586616	0.10202443	0.10028706	0.09897363	0.09890537
human	complications	georgia	constitution	nonviolence	legislation	voting
0.09883332	0.09716990	0.09696635	0.09610843	0.09607732	0.09593372	0.09576468
workers	integration	political	law			
0.09572006	0.09560082	0.09551930	0.09485102			

Summary

We introduce ***CorShrink*** as a simple and fast adaptive method for shrinking correlations and correlation matrices.

CorShrink can adjust the degree of shrinkage for each pair of variables based on the number of matched samples contributing data on both variables, thereby having the flexibility to handle large scale missing data.

Even when there is no missing observations in data, ***CorShrink*** shows competing performance with other correlation matrix estimation methods - ***corpcor, glasso*** etc - specially in n <p settings.

CorShrink allows for flexible choices of shrinkage targets and shrinkage distributions, because **ash** is flexible.

CorShrink can be used to obtain more robust/stable cosine similarity measures and hence more stable rankings between word pairs in a word2vec model fit.

Thank You !!

Any questions?

word2vec

I

w_{t-2}

$v_{w_{t-2}}$

am

w_{t-1}

$v_{w_{t-1}}$

giving

w_t

v_{w_t}

v'_{w_t}

a

w_{t+1}

$v_{w_{t+1}}$

talk

w_{t+2}

$v_{w_{t+2}}$

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

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