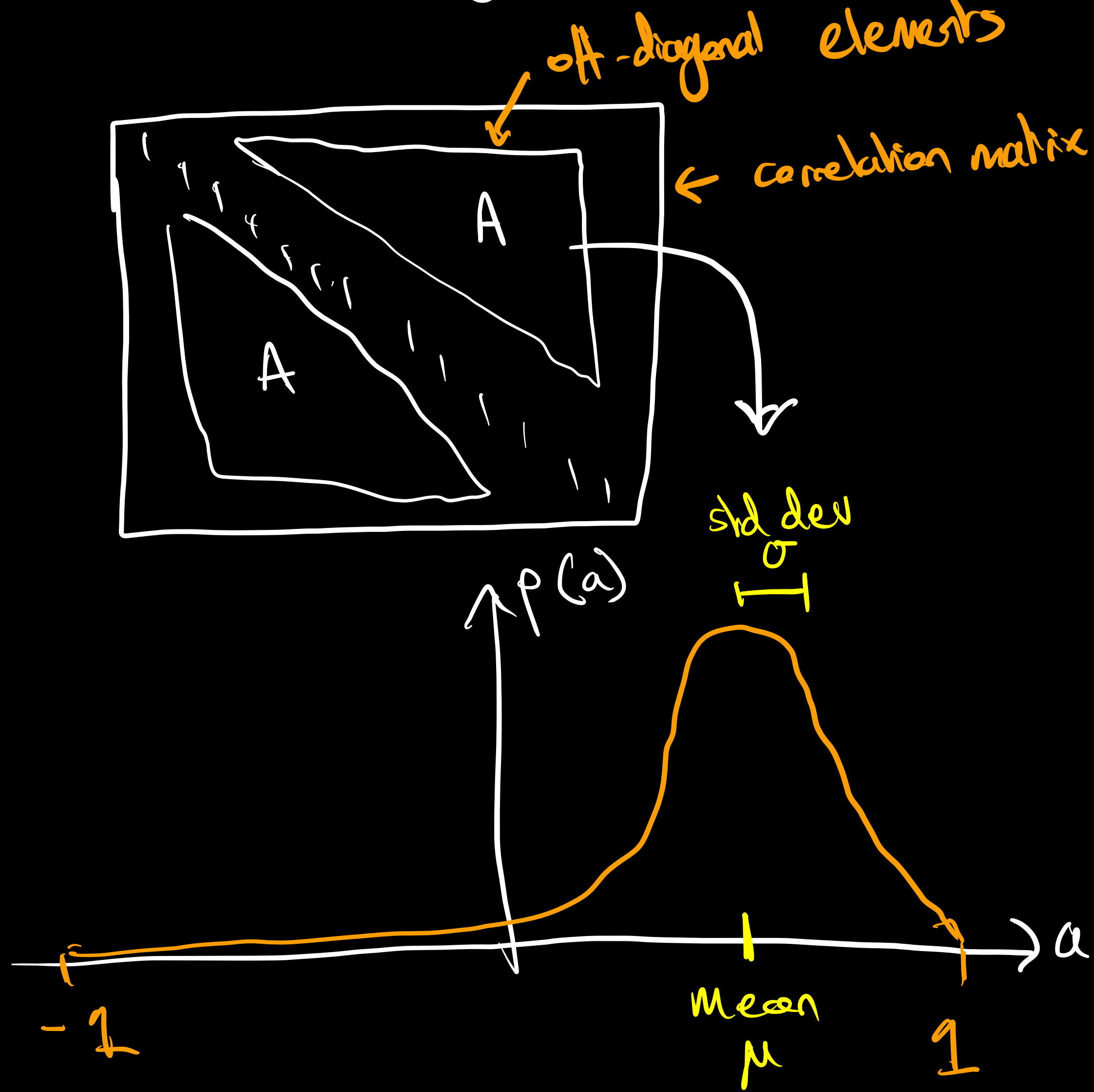
PROBLEM:

Generate random correlation matrices with particular off-diagonal distributions.



Generate correlation matrices whose off-diagonal elements have:

* mean μ
* std dev σ

Why?

Pops up occasionally, e.g. in my neuroscience research!

Why is it non-trivial?

You can't just sample random matrices, not all matrices are correlation matrices!

THERE IS DEPENDENCE BETWEEN THE ELEMENTS

If A + B are correlated high
AND B + C are correlated high

then A + C MUST be quite correlated

has to be large, can't just randomly sample

To understand this from another perspective, and for later use...

Geometric View on Problem:

Create a dataset, N datapoints, put them in matrix $\underline{\underline{X}}$

$$\underline{\underline{X}} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix} \updownarrow \text{Features}$$

← Datapoints →

$d_1, d_2, \ldots, d_N$

$\underline{a}$ ← feature vector

$\underline{b}$

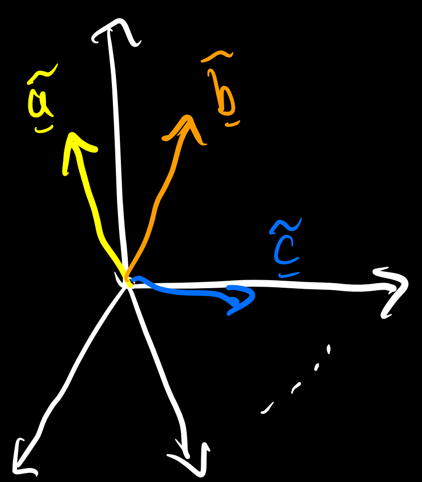
$\vdots$

$\underline{z}$

⇒ create mean-centered features

$$\tilde{\underline{a}} = \underline{a} - \mu_a \underline{1}$$
$$\tilde{\underline{b}} = \underline{b} - \mu_b \underline{1}$$
$$\vdots$$
$$\tilde{\underline{z}} = \underline{z} - \mu_z \underline{1}$$

Feature vectors are points in some N-dimensional space



$$\sigma_a^2 = \mathbb{E}\left((a - \mu_a)^2\right)$$
$$= \frac{1}{N} \sum_i \tilde{a}_i^2 = \frac{1}{N} |\tilde{\underline{a}}|^2$$

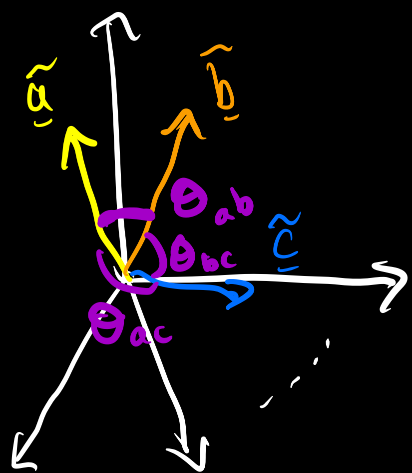
Length of the vector $\tilde{\underline{a}} = |\tilde{\underline{a}}|$ gives variance

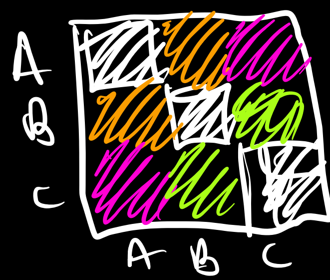
Then correlation between two features is the angle!

$$\rho_{ab} = \frac{\mathbb{E}[(a-\mu_a)(b-\mu_b)]}{\sqrt{\mathbb{E}[(a-\mu_a)^2]\,\mathbb{E}[(b-\mu_b)^2]}}$$

$$= \frac{\underline{\tilde{a}} \cdot \underline{\tilde{b}}}{\sqrt{|\underline{\tilde{a}}|^2\,|\underline{\tilde{b}}|^2}} = \cos\theta_{ab}$$

CONSTRAINT MAKES SENSE GEOMETRICALLY





$\therefore$ small θ_{ab} and small θ_{bc}

$\Rightarrow$ quite small θ_{ac}

4 Approaches to Problem

1) Onions + partial correlations

Lewandowski, Kurowicka, & Joe 2009

+ dude on internet (amoeba)

2) Via random vectors

me! Though discussed in Hardin, Garcia, Golan et al. 2013 in "A Method for generating realistic correlation matrices"

3) Via neural networks

Gautier Marti, 2019

4) Factor loadings

dude on internet

(New 2022! Via reparameterisation, Archakov et al. 2022, A new method for generating random correlation matrices)

(New 2023! Via noise addition. Hardin et al. 2013

A method for generating realistic correlation matrices)

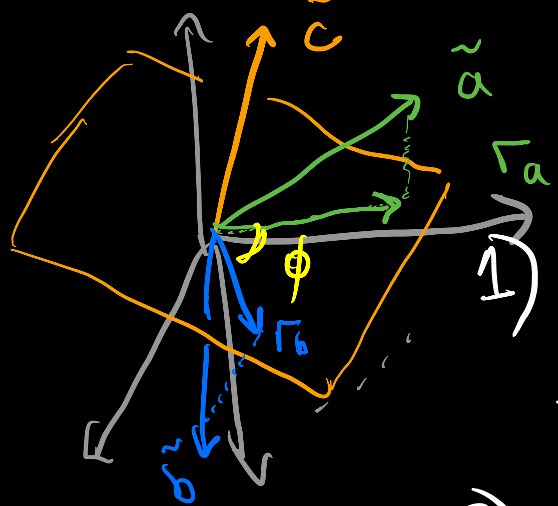
Sampling Uniformly from set of correlation Matrices

Lewandowski, Kurowicka, & Joe, 2009

VINE METHOD

key beginning: partial correlations are independent!

Partial correlation = correlation between residuals of a & b after regressing with c



Geometrically

1) Projection into plane $\perp$ to c

$\Downarrow$

residual vector after linear regression with c

2) Angle ϕ of two residual vectors

= partial correlation

$\exists$ a recursive formula for partial correlations

$$\rho_{xy\cdot\underline{z}} = \frac{\rho_{xy\cdot\underline{z}\setminus z_0} - \rho_{xz_0\cdot\underline{z}\setminus z_0}\,\rho_{z_0y\cdot\underline{z}\setminus z_0}}{\sqrt{1-\rho^2_{xz_0\cdot\underline{z}\setminus\{z_0\}}}\sqrt{1-\rho^2_{z_0y\cdot\underline{z}\setminus\{z_0\}}}}$$

$\therefore$ Can relate complete partial corr $\rho_{xy\cdot\underline{z}}$ to incomple $\rho_{A\cdot\underline{z}\setminus z_0}$ $\quad A = \begin{matrix} xy \\ xz_0 \\ yz_0 \end{matrix}$

Vine = organisation of computations

Can do a recursive computation $O(n^3)$ to get from $\frac{n(n-1)}{2}$ partials to $\frac{n(n-1)}{2}$ corrs.

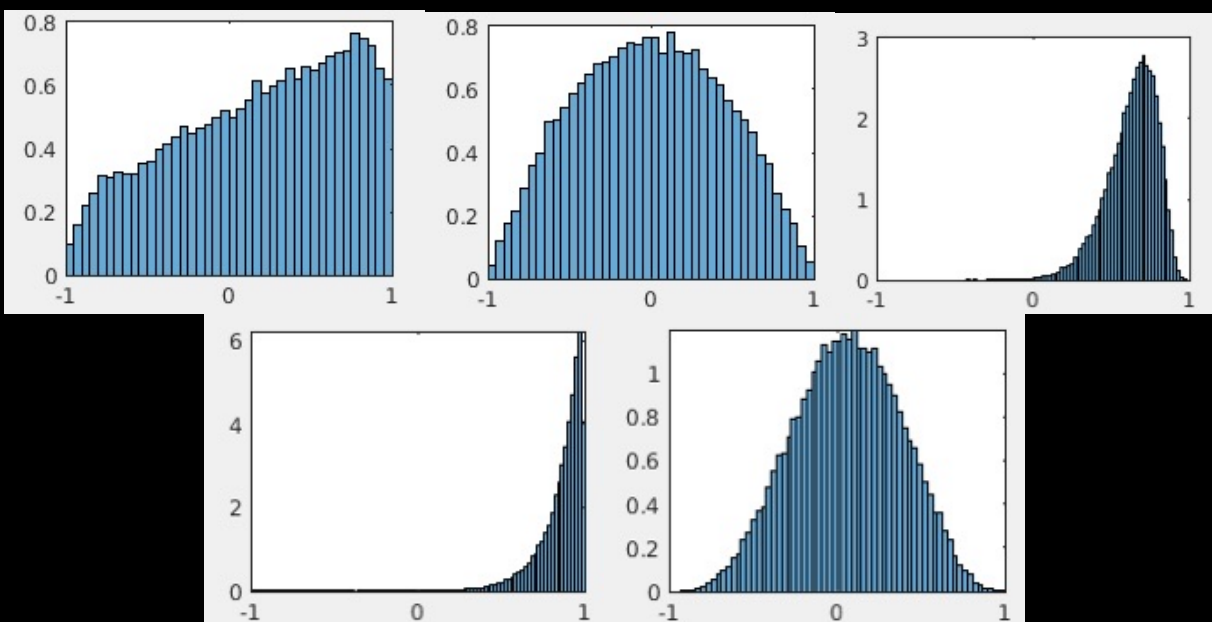
Paper shows how to sample partial correlations in order to sample uniformly from correlation matrices. OR with prob $\propto |\underline{C}|^\eta$

Alteration: man on internet 2014

$$e_{partial} \sim Beta(\alpha, \beta)$$

Vary α & β manually to get desired distribution.

Some random $\alpha + \beta$ values:



$p(e)$

e

PRO: easy to get all kinds of distributions

CON: slow recursive formula for large dimensionality

not arbitrarily complex distributions

probably hard to analyse

METHOD 2 : via random vectors

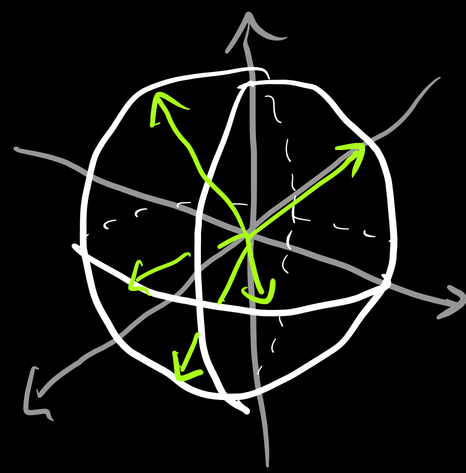
1) Generate points on sphere

- First generate

$$\underline{x} \sim \mathcal{N}(\underline{0}, \underline{\underline{1}})$$

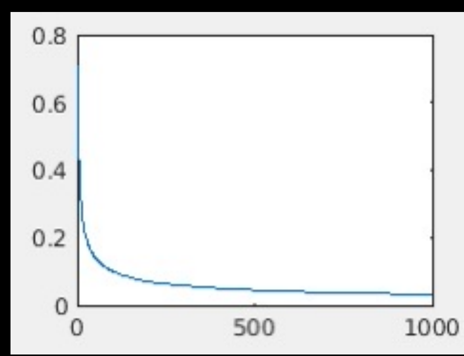
- Then normalise

$$\underline{x} = \frac{\underline{x}}{|\underline{x}|}$$



2) Get dot product matrix

How to vary dot product matrix? $\sigma(e)$



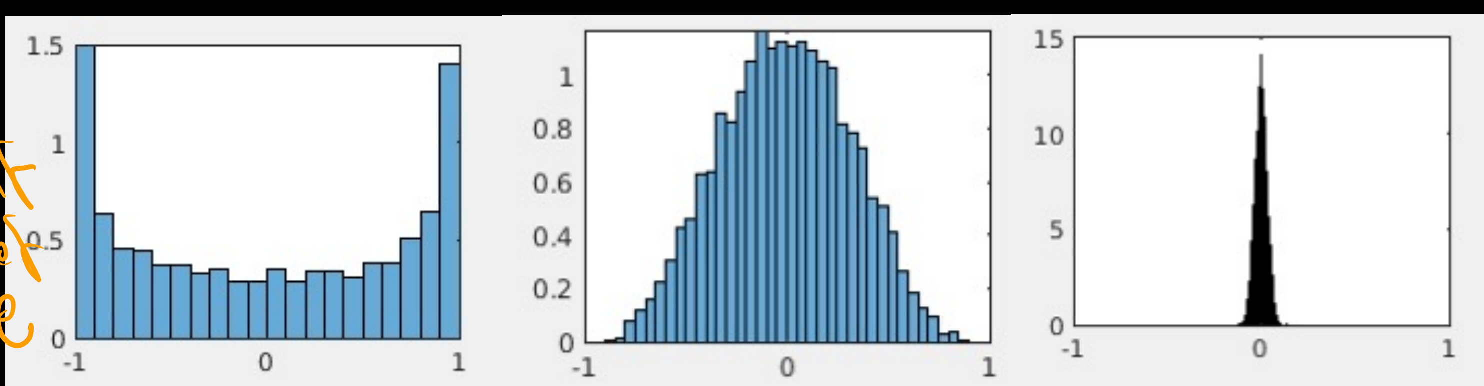
D

σ : vary dimensionality

↑ variance of dot products ↑ D

D = 2 D = 10 D = 1000

pdf of e



e e e

elements of dot product matrix

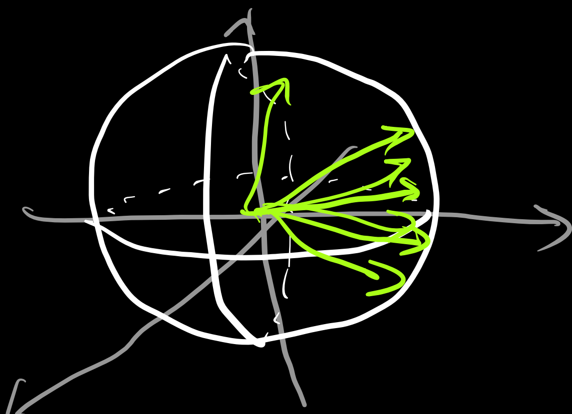
μ : shift centre

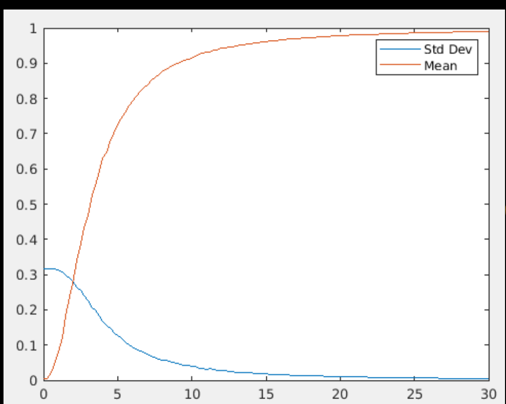
- First generate

$$\underline{x} \sim \mathcal{N}(\underline{0}, \underline{\underline{1}}) + \hat{\mu} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

- Then normalise

$$\underline{x} = \frac{\underline{x}}{|\underline{x}|}$$

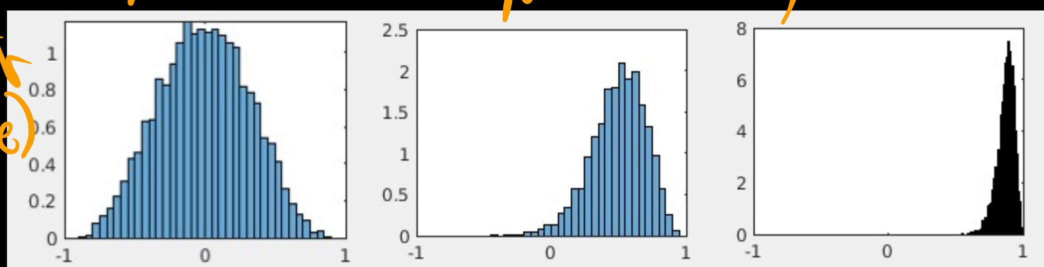




$\hat{\mu}$

$\hat{\mu} = 0$ $\quad$ $\hat{\mu} = 1$ $\quad$ $\hat{\mu} = 10$

pdf $p(e)$



e $\quad$ e $\quad$ e

Pro : very easy + quick

probably can be analysed? $\mu = f(\hat{\mu}, D)$?

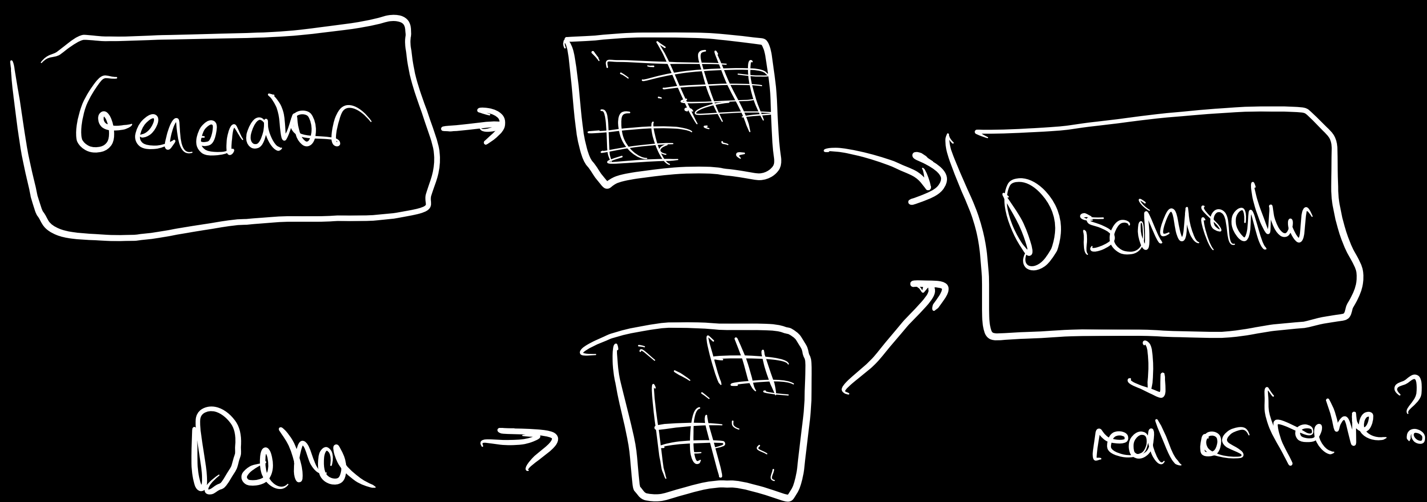
Con : have to manually tune

not arbitrary complex

METHOD 3

CORRGAN, Gautier Marti, 2019

Match measured correlation matrices
(e.g. from finance)



PRO: not just mean + variance, any structure in distribution

CON: you have to train a GAN !! :(

Method 4: Factor loadings

$$\underline{\underline{W}} \in \mathbb{R}^{k \times d}$$

$$\underline{\underline{B}} = \underline{\underline{W}}\,\underline{\underline{W}}^T + \underline{\underline{D}}$$

diagonal matrix

+ve - definite covariance matrix

$$\underline{\underline{C}} = \underline{\underline{E}}^{-1/2}\,\underline{\underline{B}}\,\underline{\underline{E}}^{1/2}$$

correlation matrix

diagonal matrix with same diagonal as $\underline{\underline{B}}$

Large k: random matrix, low off-diag corr

Small k: v. high off-diag corr

PRO: super easy + quick

probably analysable

CON: Only one degree of freedom (could introduce more)