Basic procedures for gKDR\_GMM

ver. 20230702 by Y. Iino

ver. 20230713 by Y. Iino

(newest version: <https://docs.google.com/document/d/1zXk24MmWffhk_2ey1cnBY07FRXq-3wtBwbw91vwitwg/edit?usp=sharing>)

<Important hyperparameters>

K: reduction dimension for gKDR

kGMM: number of gaussians for GMM

k (=embed\_step): invervals used for embedding (index-based)

tau (=embed\_width): number of embed\_step's used for embedding = column number in source data

direct/indirect/all/self: link types used for determining presynaptic input neurons

Even from the same data, changing these options results in different models. Therefore, these parameters are included in folder and file names.

<Other hyperparameters>

For cross validation, time series data is split into "partition\_number" (= usually 3) parts, and models are made from each.

time\_step: typically 5. Timestep intervals in which estimation is performed. Because of the limitation of data number for gKDR, every "time\_step" time points are picked up to make the gKDR-GMM models.

offset: first timepoint to be picked up. time\_step is usually 5. Therefore offsets 0, 1, 2, 3, 4 are possible. Default 0.

<Variables>

sampleID: ID number of the sample, which is under analysis. Corresponds to the file number in cleandata\_smoothened2. In this document, "samplex" stands for "sample1", "sample2", etc.

nahead: how far timesteps ahead is the target for estimation

freerun\_repeat: how many times freerun simulation is performed.

use\_salt\_input: whether sensory stimulus is applied and included in the model. Salt stimulus timing is given in a separate file, metadata/stimulation\_timing.csv

project\_folder = '/home/iino/gKDR\_GMM\_50mM/publish\_version' (example)

data\_folder = '/home/iino/gKDR\_GMM\_50mM/cleandata\_smoothened2/' (example)

Includes: 'samplex\_ratio.csv', 'samplex\_uniqNames.csv'

common\_data\_folder = [project\_folder]/common\_data

(other folders are also under project\_folder)

Above are same as before,

procedure up to 2) in "gKDR-GMM official Procedure New Eng 20230702" are supposed to be finished.

1) gKDR\_GMM\_makemodel\_cross\_K\_shift.m

Requires: [metadata\_folder]/conneurons.csv, multiconmatrix.csv, simulation\_timing.csv

[common\_data]/samplex\_data.mat

Generates:

[model\_folder]/[ model\_indirect\_k30\_tau10\_kGMM2\_cross3]/samplex\_Kx\_offsetx\_partx\_modeldata.mat

(Includes: 'uniqNames','targetcells','selicell','colscell','Bcell','gmcell','train\_span','test\_span','data','target\_train\_all','source\_train\_all'）

Generates:

[model\_folder]/[model\_indirect\_k30\_tau10\_kGMM2\_cross3]/samplex\_Kx\_offsetx\_partx\_param.mat

（Includes: 'K','autocorrthreshold','autocorrlag','link','embed\_width','embed\_step','time\_step','nahead', 'prediction\_method', 'kGMM', 'candx', 'candx2', 'eps'）

2)

gKDR\_GMM\_cross\_evaluateL\_new\_K4\_offseteach.m

Requires:

[model\_folder]/[ model\_indirect\_k30\_tau10\_kGMM2\_cross3]/samplex\_Kx\_offsetx\_partx\_modeldata.mat

[model\_folder]/[model\_indirect\_k30\_tau10\_kGMM2\_cross3]/samplex\_Kx\_offsetx\_partx\_param.mat

[common\_data]/samplex\_data.mat

[metadata\_folder]/simulation\_timing.csv

Generates:

evaluateL\_results/indirect\_k30\_tau10\_K4\_kGMM2\_new\_offsets/samplex\_offsetx.mat

(Includes: 'p\_array','targetcellnames','sumlogL0\_array','permute\_salt')

3)

gKDR\_GMM\_connection\_strength\_indirect\_cross\_K\_offset.m

Requires:

[model\_folder]/[ model\_indirect\_k30\_tau10\_kGMM2\_cross3]/samplex\_Kx\_offsetx\_partx\_modeldata.mat

[model\_folder]/[model\_indirect\_k30\_tau10\_kGMM2\_cross3]/samplex\_Kx\_offsetx\_partx\_param.mat

Generates:

[models\_folder]/model\_indirect\_k30\_tau10\_kGMM2\_cross3/samplex\_Kx\_offsetx\_partx\_gradient.mat

4)

evaluate\_connection\_strength\_cross\_offsets.m

(on HGC: qsub -t 1:24 -pe def\_slot 1 got30.sh evaluate\_connection\_strength\_cross\_offsets)

Requires:

[common\_cell\_order\_folder]/samplex\_common\_outperm.mat

[models\_folder]/model\_indirect\_k30\_tau10\_kGMM2\_cross3/samplex\_Kx\_offsetx\_partx\_gradient.mat

Generates:

[connection\_strength\_folder]/indirect\_cross3/samplex\_Kx\_kGMM2\_offsetx\_partx\_weightmatrix.mat

(Optionally)

[connection\_strength\_folder]/indirect\_cross3/samplex\_Kx\_kGMM2\_offsetx\_partx\_timeweight\_cellx\_indirect.tif/fig, \_weightmatrix\_indirect\_Kx\_samplex.tif/fig, \_All\_network.tif/fig/pdf(not optional)

5)

process\_weight\_for\_allnetwork.m

Requires:

[connection\_strength\_folder]/indirect\_cross3/samplex\_Kx\_kGMM2\_offsetx\_partx\_weightmatrix.mat

Generates:

[connection\_strength\_folder]/indirect\_cross3max\_all(direct)\_link\_FDR0.005/process\_weight\_allnetwork\_core\_phys\_offsets.mat, validglobaltable.tif/fig/csv

Includes: globalvalidweightarray and globalsigweightnonself(the latter is not used)

6)

draw\_connection\_gridmaps\_eachsample.m

Requires:

[connection\_strength\_folder]/indirect\_cross3max\_all(direct)\_link\_FDR0.005/process\_weight\_allnetwork\_core\_phys\_offsets.mat

Generates:

[connection\_strength\_folder]/sample1-24\_pcolor\_single.ps/pdf　：log10(1/p)のグリッドマップ

[connection\_strength\_folder]/sample1-24\_brickwork\_single.ps/pdf

[connection\_strength\_folder]/sample1-24\_brickwork\_samplex.tif/fig/pdf　：煉瓦積みのグリッドマップ

[connection\_strength\_folder]/sample1-24\_lagweights\_single.ps/pdf

[connection\_strength\_folder]/sample1-24\_lagweights \_samplex.tif/fig/pdf　：時間遅れ重みのサンプルごとのマップ

7)

draw\_connection\_gridmaps\_allsamples.m

Requires:

[connection\_strength\_folder]/indirect\_cross3max\_all(direct)\_link\_FDR0.005/process\_weight\_allnetwork\_core\_phys\_offsets.mat

Generates:

[connection\_strength\_folder]/allsamples\_pcolor\_single.ps/pdf

[connection\_strength\_folder]/allsamples\_brickwork\_single.ps/pdf

[connection\_strength\_folder]/allsamples\_brickwork\_samplex.tif/fig/pdf

[connection\_strength\_folder]/allsamples\_lagweights\_single.ps/pdf

[connection\_strength\_folder]/allsamples\_lagweights \_samplex.tif/fig/pdf

(Note)

5)のprocess\_weight\_for\_allnetwork.mがやっていること。

１）各sample, K, offset, partitionについてweightmatrixを読み込みweightarray(Mt+1,Mt,partition\_number,length(offsets),length(Ks))にまとめる。

２）sampleのp\_array(bootstrap testの結果)を読み込む。

３）各配列を計算。

p\_array\_max\_valid(Mt,partition\_number,length(offsets),length(Ks))はp\_arrayによるcross validationで他の二つのpartitionの両方のpがalpha (現在0.001)より小さいものを示す論理配列。

validweightarrayはweightarrayのうち、モデルが上のテストをクリアしているものだけ（ただし、partition, offset, Kの45組み合わせのうち5つ以上クリアしたもののみ）。

つまり、validがついているのはモデルがvalidifyしたという意味。

次に、validweightarrayの(row,col, :, :,:)（through partition, offset, K）にranksumテスト,FDRを施してFDRthre以上でweightのプラスマイナスが安定している神経ペアだけのweightとしたもの（これ以外はNaN; 自分から自分もNaN）がsigweightnonself。ただし、only\_directとした場合は直接結合以外はfalse（=0）としている。

つまりsigがついているのはranksumテストを通ったということ。ただし、sigweighnonselfは今後ほとんど使われておらず、改めてranksumテストを行なっている。

４）global参照としてglobalweightarray(Mg+1,Mg,partition\_number,length(offsets),length(Ks),length(samples))にweightarrayをそのまま入れる。

globalvalidweightarray(Mg+1,Mg,partition\_number,length(offsets),length(Ks),length(samples)= 160,159,3,5,3,24)にvalidweightarrayをそのまま入れる。

globalsigweightnonself(Mg+1,Mg,length(samples))にsigweightnonselfをそのまま入れる。

connection\_strength/indirect\_cross3max\_all\_link（または\_direct\_link） \_FDR0.005/process\_weight\_allnetwork\_core\_phys\_offsets.mat

に保存。