Basic procedures for gKDR\_GMM

ver. 20230702 by Y. Iino

<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)

0) preparation

run normalize2 in test200219\_rica\_tcrs4d.m

This applies median filter smoothing, removes abnormal values, detrend and normalize.

Then, run mat2csv.m to make samplex\_ratio.csv, samplex\_uniqNames.csv etc.

For connectome data, we use synapse\_All\_Iwasaki2.txt in preparation\_codes folder, which was provided by Yuishi Iwasaki from Ibaraki University, and the data is from CCEP project and WormAtlas data, both digitalization of John White's connectome paper.

http://ims.dse.ibaraki.ac.jp/ccep/

Use this data and run make\_connectivity\_matrix.m to make conneurons.csv and multiconmatrix.csv among others.

1)

Run code: make\_common\_data.m

Hyperparameter: autocorrthreshold = 0.3, autocorrlag = 20

Requires: 'conneurons.csv', 'samplex\_ratio.csv', 'samplex\_uniqNames.csv'

Generates: [common\_data\_folder]/samplex\_data.mat

Includes: 'data', 'uniqNames', 'targetcellnames', 'targetcells', 'Mt', 'autocorrthreshold', 'autocorrlag'

Next, run: common\_clustering.m

Requires: ‘conneurons.csv’, [common\_data\_folder]/samplex\_data.mat

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

2)

common\_clustering.m

Requires: '../conneurons.csv',

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

Includes: 'outperm'

Generates: [common\_cell\_order\_folder]/samplex\_corrmatrix.tif/fig

3)

gKDR\_GMM\_makemodel.m

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

Generates: [model\_folder]/samplex\_Kx\_modeldata.mat

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

Generates: [model\_folder]/samplex\_Kx\_param.mat

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

4)

gKDR\_GMM\_freerun\_indirect\_K234\_kGMM3.m

Generates:

[simulation\_result\_folder]/indirect\_kx\_taux\_Kx\_kGMMx/samplex/samplex\_freerunx\_savedata.mat

Includes:

'uniqNames','targetcells','colscell','sampleID','targetcellnames','Bcell','gmcell','saltdata', 'selicell','Tc\_real','Tc\_all','real\_predict', 'outperm', 'rmatrix','train\_span','target\_train\_all','source\_train\_all'

Generates (optional):

samplex\_real\_heatmap.fig/.tif

samplex\_real\_corrmatrix.fig/.tif

samplex\_freerunx\_heatmap.fig/.tif

samplex\_freerunx\_corrmatrix.fig/.tif

(gKDR\_GMM\_freerun\_indirect\_K234\_kGMM3\_sweeper.m; which redoes failed samples)