Epilepsy study combined with Granule and mossy cell regenerated stem cell sequences and clustering monitored by electrode data with Bayes Inference Model

Qin He[a,b], Jarri Hyttinen[a], Sampsa Pursiainen[a], Jukka Peltola[a,b]
a. Tampere Universities b. TAYs(University Hospital)
qin.he@tut.fi/qin.he@tuni.fi

Abstract

The two typical excitatory cell types granule cell(GCs) and mossy cells of DG(dentate gyrus) see fig 1 in chronically epileptic mice identified by optogenetic tagging are studied through the electrographic data of the seizure. For both GC and MC, the reconfiguration are tested related to spatial representations via the emergence, extinction and rate alteration of firing fields. To investigate the consequences of MC and GC loss on cognition, the activities of the cell neurons are recorded by MEA/EEG/MEG. Further, gene expression(fig:2) and protein analysis is applied to understand the functionality of them in addition to the intrinsic level. One machine learning method based on Bayes Variational Inference model is applied to the t-SNE[1] clustering of the genes. As DNA and RNA are usually of high dimension, the Bayes variance is utilised to modify the BIC in guaranteeing the convergence, computing KL and perplex and the cluster distance are measured with Euclid metric simulating with normal and gamma distribution. In the MEA analysis, another machine learning method combining Hamilton Markov chain likekihood-free model[2, 3] is applied on ictal classifications as well as different auto-correlation, gaps, asymmetry, amplitude, phase angle and etc. on different time windows and bands. In the genetic analysis, more analysis with the expression is on regenerated data according to its distribution. One review comparing epilepsy with AD in genetic regulation is introduced as well for giving guide on epilepsy gene detection method as well. More prominent work see the references[12].

Model Formulation

The Bayes Variational Inference[9] we use here is to minimize the Kullback-Leiber(KL)[10] divergence, the log difference between observed and approximated posterior which can also be regarded as an optimization problem: Modifying the t-Distributed Stochastic Neighbor Embedding(t-SNE), the pairwise distances in high dimensional space with data points x_i is converted to corresponding embedding points y_i pairwise join distributions in low dimensional spaces, which respectively follows:

$$q_{i,j} = \frac{(1+|y_i-y_j|^2)^{-1}}{\sum (1+|y_s-y_t|^2)^{-1}}$$

while high dimensional one is defined in symmetrical conditions:

$$p_{i,j} = (p_{i|j} + p_{j|i})/2n$$

,where

$$p_{i|j} = \frac{exp(-|x_i-x_j|^2/2\sigma_j^2)}{\sum exp(-|x_s-x_t|^2/2\sigma_j^2)}$$

and the KL to be optimised is thus:

$$KL(P|Q) = \sum p_{i,j} * log\frac{p_{i,j}}{q_{i,j}}$$

Note that the σ_j is optimized through bisectional search automatically with the pre-saticified perplexity

$$Perplex(p_j) = 2^{H(P_j)}, \text{ where } H(P_j) = -\sigma_j p_{i|j} * log2^{p_{t|j}}$$

so that

$$Perplex(p_j) = Perplex$$

, where Perplex is the hyperparameter of the t-SNE central to the final cluster.
Large Perplex usually leads to the embedding sub-optimal in detecting the pattern of the data(In the limit, when the Perplex goes to the number of data points, the corresponding embedding form a Gaussian or uniform like distribution and fails to be useful for structure detection at all) and thus, we design a new criteria:

$$S(Perplex) = KL(P|Q) + log(n) * \frac{Perplex}{n}$$

with inspiration of the Bayesian Information Criteria(BIC):

$$BIC = -2 * log(L) + log(n) * k$$

, where the first term stands for the goodness-of-fit of the maximum-likelihood-estimation and the second controls the complexity of the model with penalty k scaled by log(n).

Performance in Clustering with Bayes Variational Inferenced t-SNE

Intuitively, when Perplex increases, differences among points will become less and less significant with regard to the length of the kernel in distribution P, and P will tend to uniform.The forward form of KL has large cost for under-estimating probability but not for over-estimating. That is, if $p_{i,j}$ is large and $q_{i,j}$ is small, KL divergence is large while in the opposite direction, KL is not affected. Increasing Perplex leads to larger σ_j and more uniform $p_{i,j}$ so it is easier is for the student-t distribution in low dimensional space to assign mass for all probability points sufficiently. This is the so called crowding problem: When projecting from high to low dimensional space, there is not enough room in lower dimensional space.Generally, increasing Perplex relaxes the problem and reduces the amount of structure to be modelled with less error according to KL while pays a cost in the second term. With the practical test, we apply it on the MC GC cell expression classification(fig 3 - 5) with t-SNE.
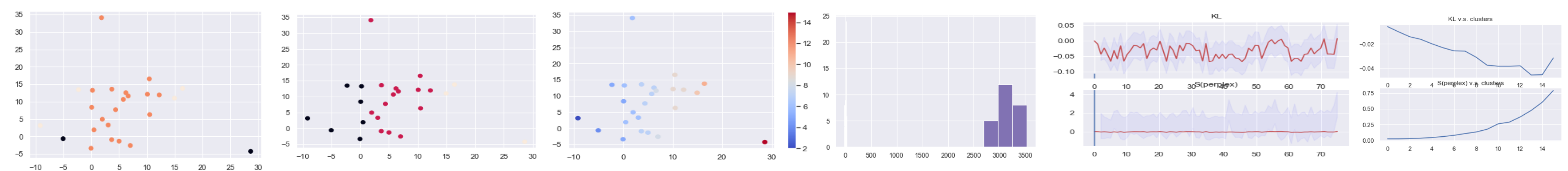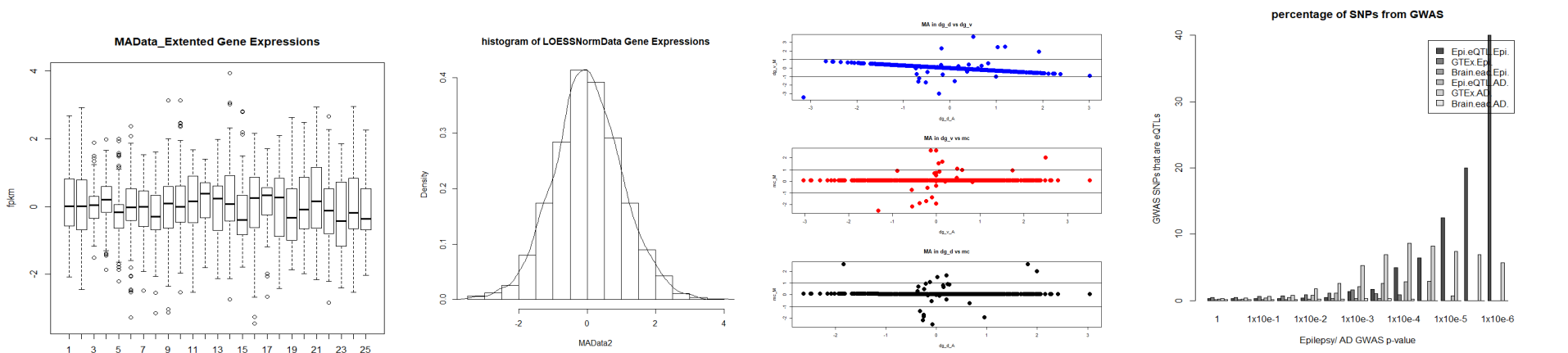

MEA Analysis

There are 5 ictal periods recorded (10 bins of interested separately). For the whole dataset, the auto-correlation is calculated to detect the onset of seizure and preprocess including baseline correction and normalization after filtering are applied. As the seizure is usually recorded on high frequency band compared to other signals, for instance the signal for learning tasks, the MEA signal is studied comparing different bands as well as on the whole 180 s time period. According to the fig 9, 11 , the signal is oscillated evenly on the gamma band (25 Hz to 100 Hz) and sparsely detected as spikes at delta, theta, alpha, mu, SMR and beta bands according to the fpkm values (on log level) of the genes in the range -3 to 4. More results about train and bursts detection as well as auto-correlation(fig:10) and phase analysis of each firing spikes and its types mainly based on duration and asymmetry(fig:12) see [17-19].

Genetic analysis

As study of cellular processes existing inside intact organisms requires methods to visualize cellular functions such as gene expression in deep tissues, the 75 regenerated expression dataset (seen in fig 13, dg-d, dg-v and MC cell separately)is studied along with the gene regulation as the substrate of the observable trait at which the genotype gives rise to the phenotype (further cellular differentiation adaptability of organism in the versatility and adaptability).After the LOESS normalization, it is normalized according to QC test(fig:14) and most genes of mc and dg-d are expressed stationarily with MA value(fig:15) mainly distributed around zero (since the expression of genes are assumed to be unchanged at certain tissue region.)while has decreasing trend for the dg-v expression.(More in log2 expression and MDR plot can be found in ref [12], including one review about epilepsy and AD comparison study through eQTL, fig: 16.)

Ictal classification with Bayes Inference

$$\lambda \leftarrow (\log \hat{r}(\mathbf{x}|\boldsymbol{\theta}_k) + \log p(\boldsymbol{\theta}_k)) - (\log \hat{r}(\mathbf{x}|\boldsymbol{\theta}_t) + \log p(\boldsymbol{\theta}_t)) + K(\mathbf{m}_k) - K(\mathbf{m}_t)$$

$$\rho \leftarrow \min(\exp(\lambda), 1) \qquad \boldsymbol{\theta}_{t+1} \leftarrow \begin{cases} \boldsymbol{\theta}_k & \text{with probability } \rho \\ \boldsymbol{\theta}_t & \text{with probability } 1-\rho \end{cases}$$

Bayes inference model[5,15] combined with Hamilton Markov(fig: 17)[2] sampling is utilized as the updating of dynamic information of the MEA signal, with the prior knowledge of the different ictal patterns decomposed by empirical mode decomposition(EMD) with intrinsic mode functions(IMF), after computing the entropy, instantaneous energy and frequencies, it is classified[14] based on likelyhood-free method. To further study the ictal/interictal, HFO parts are extracted with filter threshold as 80Hz(ripples) and 250Hz(fast ripples) separately and with more than 9 spikes detected in bins of duration at least 10ms.

Conclusions

In summary, according to the correlation is largest within one window on delta band and gamma band (with the maximum around 90 ms. The histogram and the gap can be seen distributed with gamma and normal distribution. According to the filtered signal, the type of the spike is classified as ' broad, broad, broad, narrow, broad, narrow, broad, broad, broad, broad' with (fig 18)linear regression model(acc:0.9 fig: 19): T ype`asymmetry+duration+Asymmetry/distance+Q3Distance(gap)+Q3Distance/sigma and finally into ' interictal, ictal, ictal, interictal,interictal' after (fig: 20)IMF(acc=0.8 in fig: 21). Detailed sensitivity, specificity, precision, missrate and false discovery and other evaluations along with ANOVA tests can be found in [12](bin 80 see fig 22)

With the normalized expression data (p = 0.0035 with ks-test), Genes expressed in the 3-sigma distance, Cdc45, Prox1, St18, Marsckl1, SOX2, Cdk4 and etc. known to have distinct functions are shown in the heatmap and accordingly posses correlation with close to zero in the correlation figure. However, to explore the difference of the function on cellular scale, the welch two sampled t-test is applied. According to the result, the gene expression for dg-d, dg-v and mc(p = 0.009724, p = 0.008046 and p = 0.008584) are not of the same mean on log2 level. Generally, dg-d and mc are expressed more evenly while dg-v is distributed on some specific dimension (correction method will be applied in the next step exploration.) To explore the transcription of DNA in the future, we also introduced one review of genetic regulation of gene expression in human hippo-campus [13].

References

[1] G.E van der Maaten, L.J.P.; Hinton. *Visualizing Data Using t-SNE*. Journal of Machine Learning Research, 2008.

[2] LouppeUniversity of Liège Joeri Hermans, Volodimir Begy Gilles. *Likelihood-free MCMC with Amortized Approximate Likelihood Ratios*.

[3] Annika Aktiainen Qin He, Jarri Hyttinen. *Classification of epilepsy ictal with Bayes inference Model*. Neuroelectronic Interfaces (GRS) Gordon Research Semina, 2019.

[4] Dentate granule and mossy cells exhibit distinct spatiotemporal responses to local change in a onedimensional landscape of visualtactile cues.

[5] *Lévesque Robustness Guarantees for Bayesian Inference with Gaussian Processes*.

[6] Video image-based analysis of single human induced pluripotent stem cell d erived cardiomyocyte beating dynamics using digital image correlation.

[7] *Distinguishing Hidden Markov Chains*.

[8] Video image-based analysis of single human induced pluripotent stem cell d erived cardiomyocyte beating dynamics using digital image correlation.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Estimating the dimension of a model*. Annals of Statistics, 2016.

[10] S Kullback. *Information Theory and Statistics*. John Wiley Sons, 1978.

[11] *On the Importance of Strong Baselines in Bayesian Deep Learning*.

[12] *RNA single sequences report*.

[13] *Genetic regulation of gene expression in the epileptic human hippocampus*.

[14] *Two different interictal spike patterns anticipate ictal activity in vitro*.

[15] A. Koulouri V. Rimpilainen C.H. Wolters S. Pursiainen A. Rezaei, M. Antonakakis M. C. Piastra Q. He. *Conditionally Gaussian Hierarchical Bayesian Inversion of the P50-N80 Component of SEP/SEF measurements with FEM Forward Modeling*.

[16] (Brainstorm) EEG and epilepsy.

[17] (FieldTrip) Processing and analysis of spike-train data.

[18] (SPM12) fMRI and some clinical practical data processing.

[19] Interictal high-frequency oscillations (HFOs) as predictors of high frequency and conventional seizure onset zones.