Topological Analysis of Biosensing Signal Based on Low-Voltage Alternate Current Electrokinetics

Zhifei Zhang

University of Tennessee, Knoxville

zzhang61@vols.utk.edu

January 21, 2015

Outline

Alternate Current Electrokinetics

ACEK

An inhomogeneous AC electric field is applied through microelectrodes to an aqueous solution in biosensing.

Time

After 2010, ACEK began to be wildly applied in biosensing.

Pro

Shorter assay time and/or higher sensitivity.

Con

High voltage may cause electrochemical reaction; low voltage makes biosensing signal indistinguishable.

Why signal processing is needed?

Low-voltage ACEK

We prefer lower voltage that is safe but inefficient. And the development of hardware is always lags behind software.

- signals mix up under lower voltage
- Resolution is limited by hardware
- signal processing potentially increases the performance

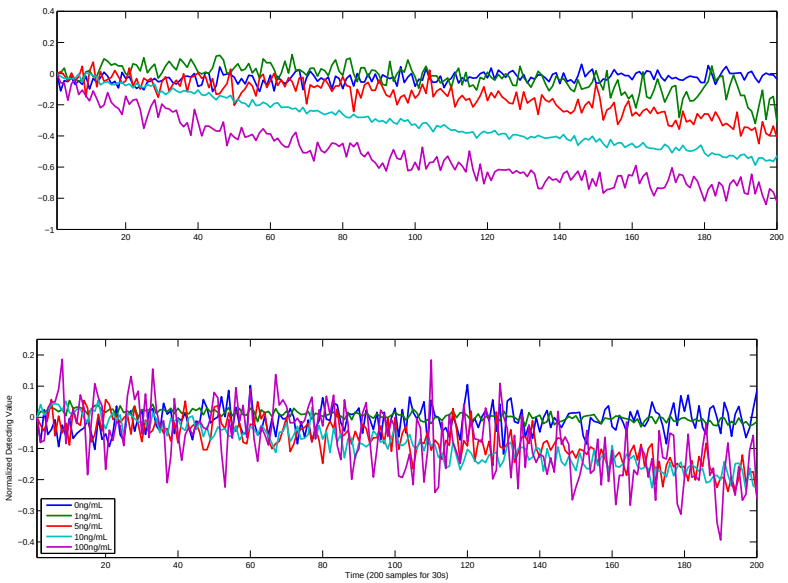


Figure: Signals of higher (top) and lower (bottom) voltage

Related work and our goal

Existing work:

- need voltage around 10V
- process periodical signal
- the signals have certain distinguishable trend

Our goal:

- need voltage below 135mV
- process non-periodical signal
- the signal have large overlap and uncertainty

Challenges

Overlap

Signals mix up together

Random oscillation

Neither frequency nor amplitude is fixed

Uncertainty

Signals of the same sample are significantly different

Limited data

Small data size

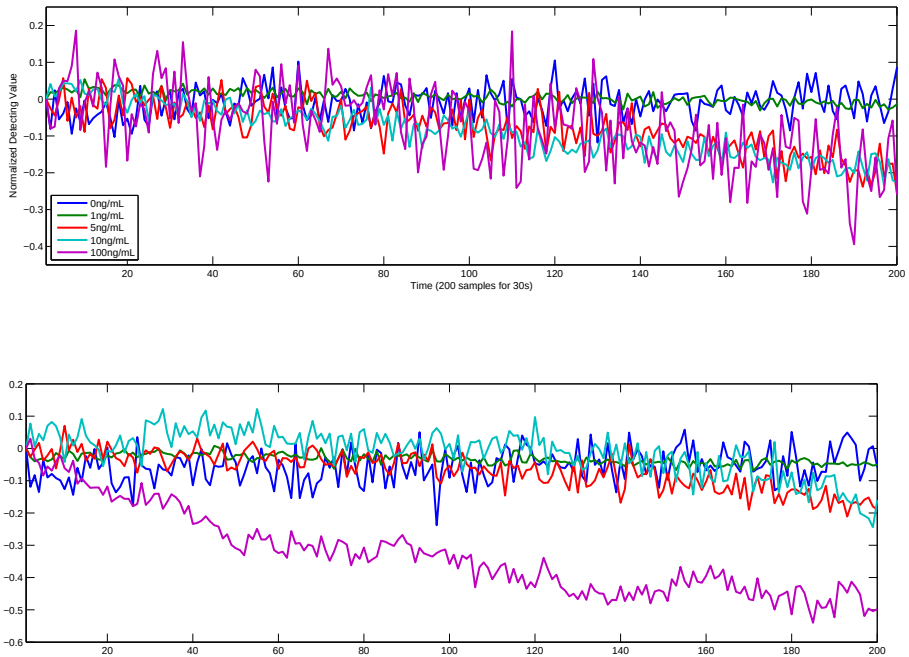
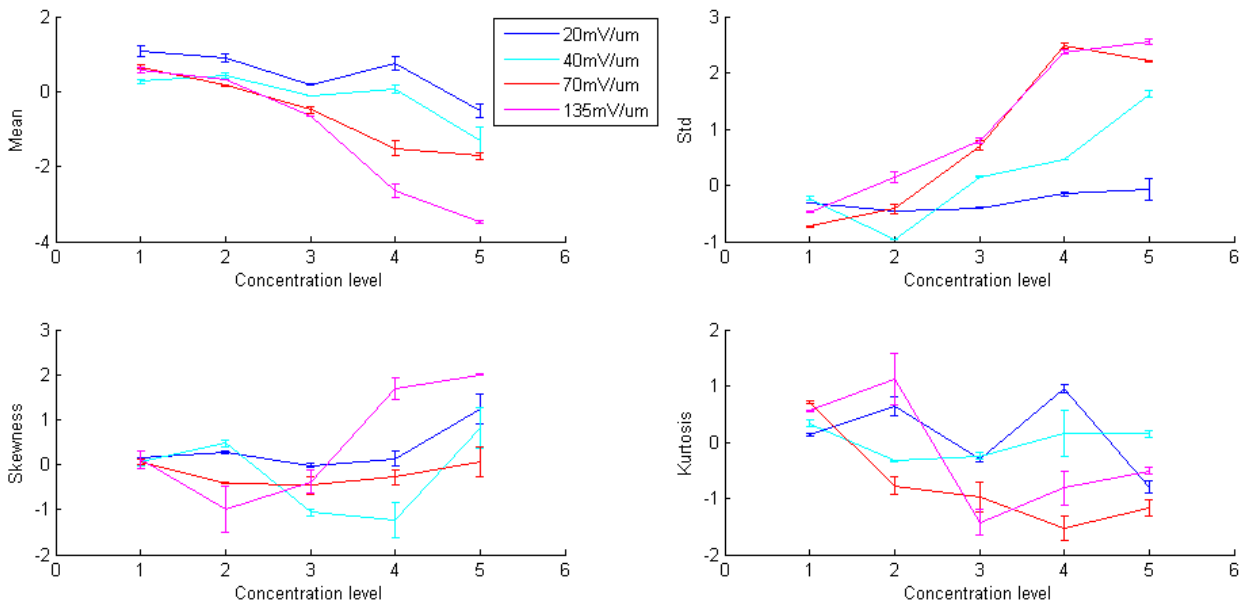


Figure: Signals under 40mV (top: tests on different concentrations; bottom: repeat the tests)

Statistical Methods

Non-monotonic mapping

Since overlap and uncertainty, statistical metrics like mean, variance, skewness and kurtosis cannot appear a monotonic variation as the concentration increase.

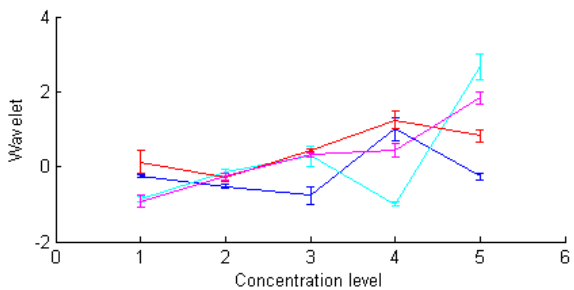


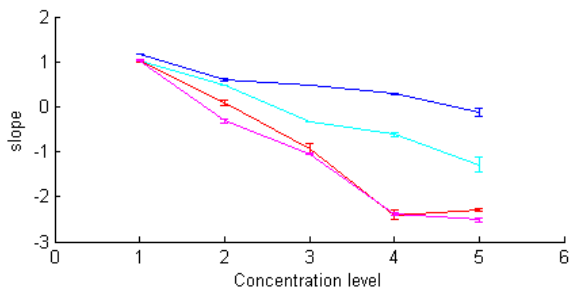
Frequency and Modeling

Fail to handle uncertainty

In the frequency domain, no fixed pattern can be extracted because of random oscillation.

The slop obtained by line fitting gives some hope, which is destroyed by the uncertainty.

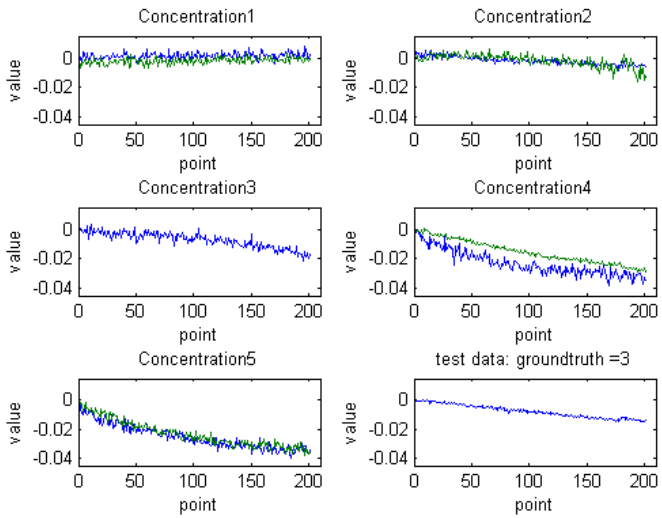




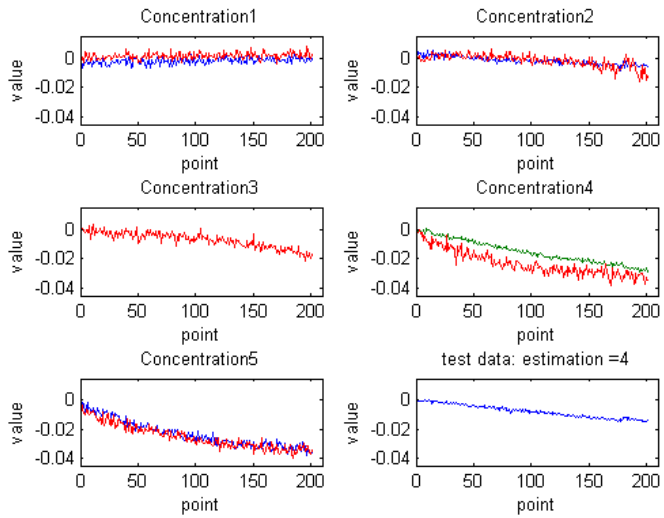
Unmixing Method

Confused by uncertainty

Uncertainty makes unmixing method confused. Moreover, the small data size makes this method meaningless.



(a) Raw data of different concentration. Bottom right: leave one out for test



(b) Class-wise endmembers. Bottom right: estimated result

Motivation

Transfer to higher dimensional space

In time and frequency space, it is hard to find a strong feature that tightly correlates to the ground truth.

Highly fused features

In topology, some features can reflect multiple features in time and frequency domain. Somewhat like feature fusion.

Potential of extenting to general signal processing

For periodical or monotonic signal, topological method should yield better performance. Thus, a general framework can be established.

Delay Embedding

Suppose a signal sequent can be presented by a discrete function $f(t),\ t \in \mathbb{Z}^+$ corresponding to the location of a sampling point. Choose a delay step $s \in \mathbb{Z}^+$ and a target dimension $D \in \mathbb{Z}^+$. The time delay embedding of f at t can be expressed as

$$DE_{s,D}f(t) = \begin{bmatrix} f(t) \\ f(t+s) \\ \vdots \\ f\left(t+(D-1)s\right) \end{bmatrix}$$

Point Cloud

Assume there are n samples in a signal sequence. If each sample is applied by the DE, $m = n - (D-1)s + 1$ vectors will be obtained finally. Those D dimensional vectors are call point cloud $\mathbf{C}$ that is written as

$$\mathbf{C} = \{DE_{s,D}f(t_1), DE_{s,D}f(t_2), \cdots, DE_{s,D}f(t_m)\}$$

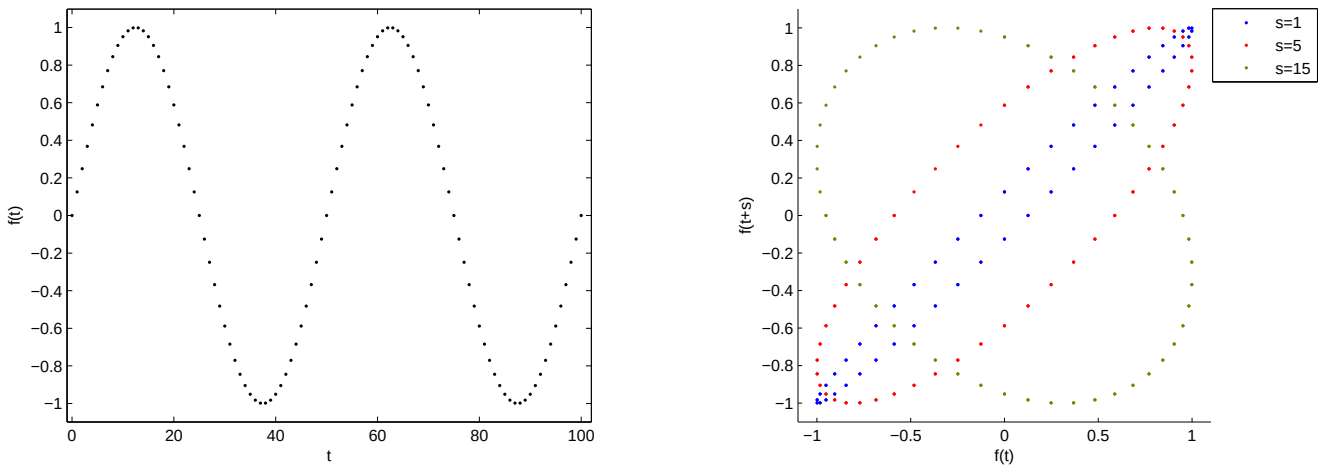
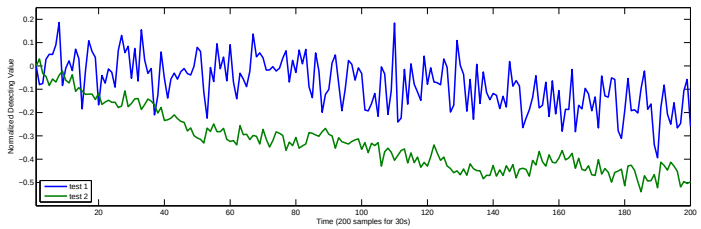


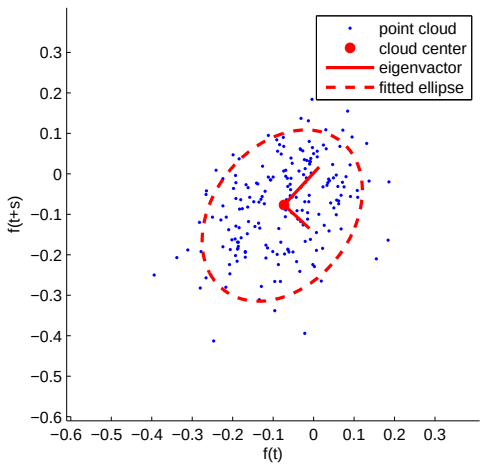
Figure: Left: $f(t) = \cos(Lt)$. Right: point cloud of different s, $D = 2$

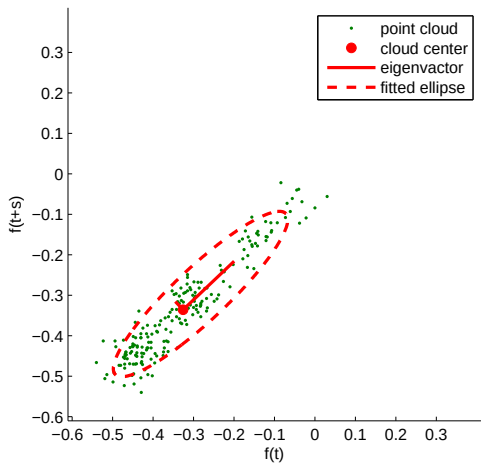
Shape Analysis

Non-periodical signal

We cannot obtain such a beautiful point cloud through delay embedding on a non-periodical signal. However, the shape of point cloud can still reflect certain features of the signal.







Shape Analysis

Shape features that free from overlap and uncertanty

Center, orientation, axis length and ration, volume, etc.

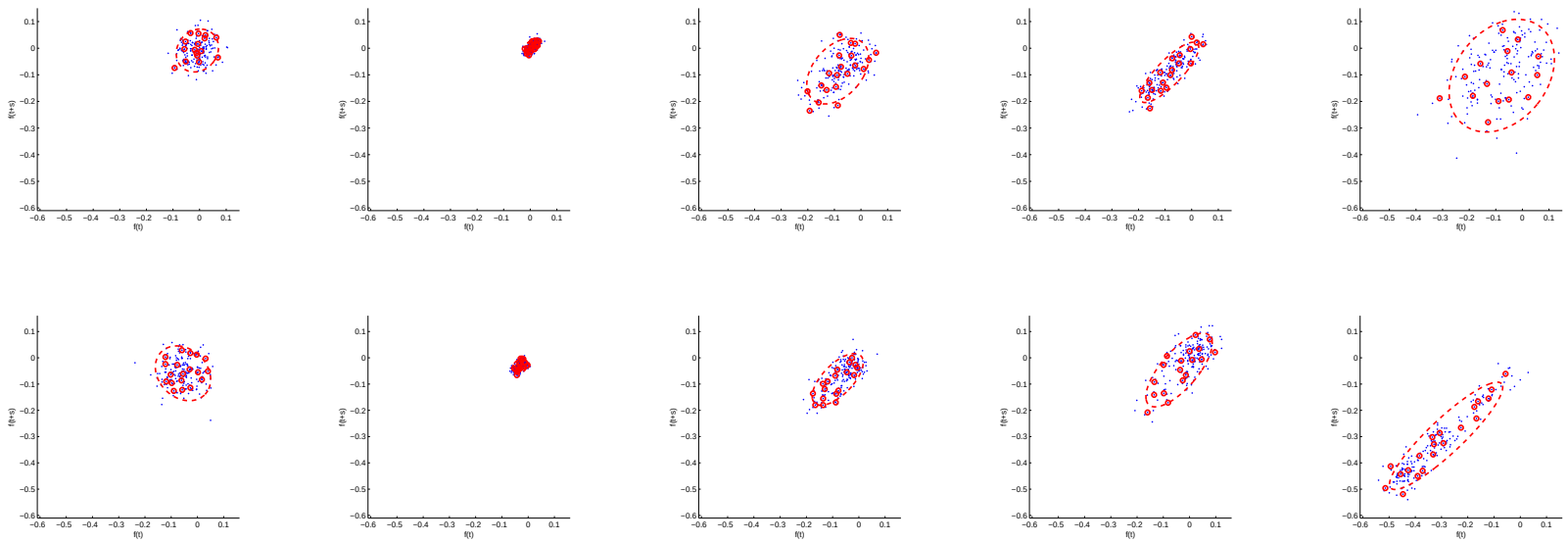


Figure: Point could under 40mV. Columns: different concentration in increasing order. Rows: two repeated tests.

Landmark Selection

Landmarks

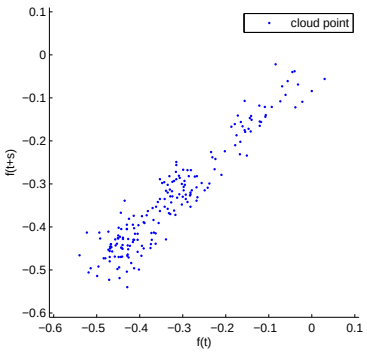
Speed up computation and illustrate inner structure of the point cloud.

Mean-shift

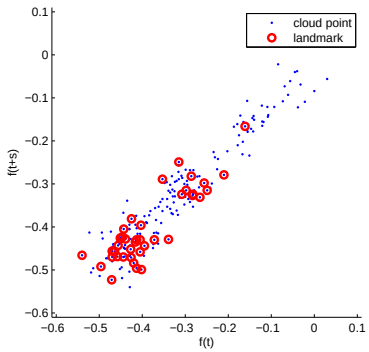
Locate dense areas.

Acnode-reduce

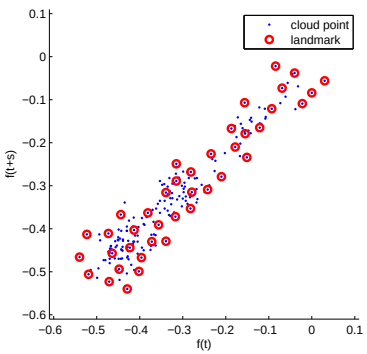
Kick out outliers.



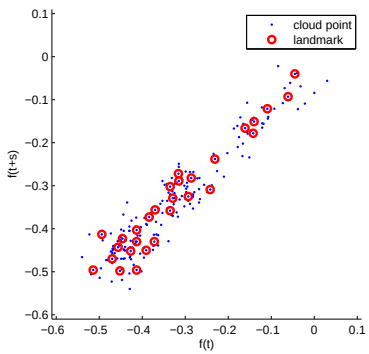
(a) Original point cloud



(b) Random

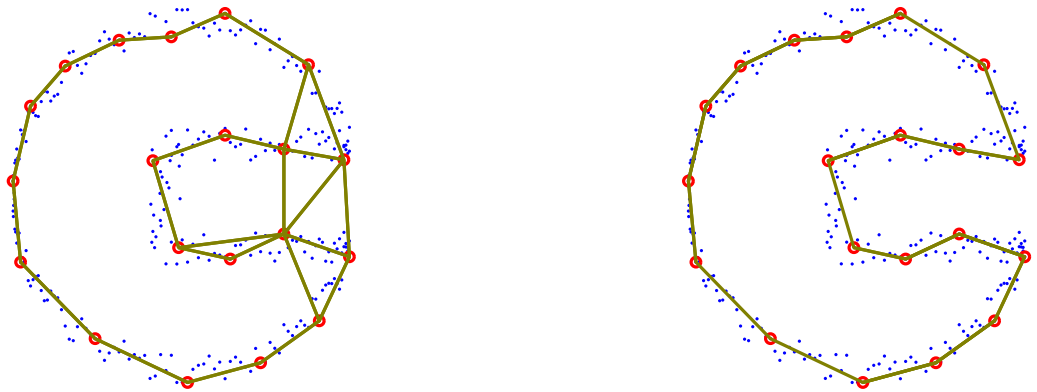


(c) Maxmin



(d) Mean-shift & acnode-reduce

Witness Rips Complex

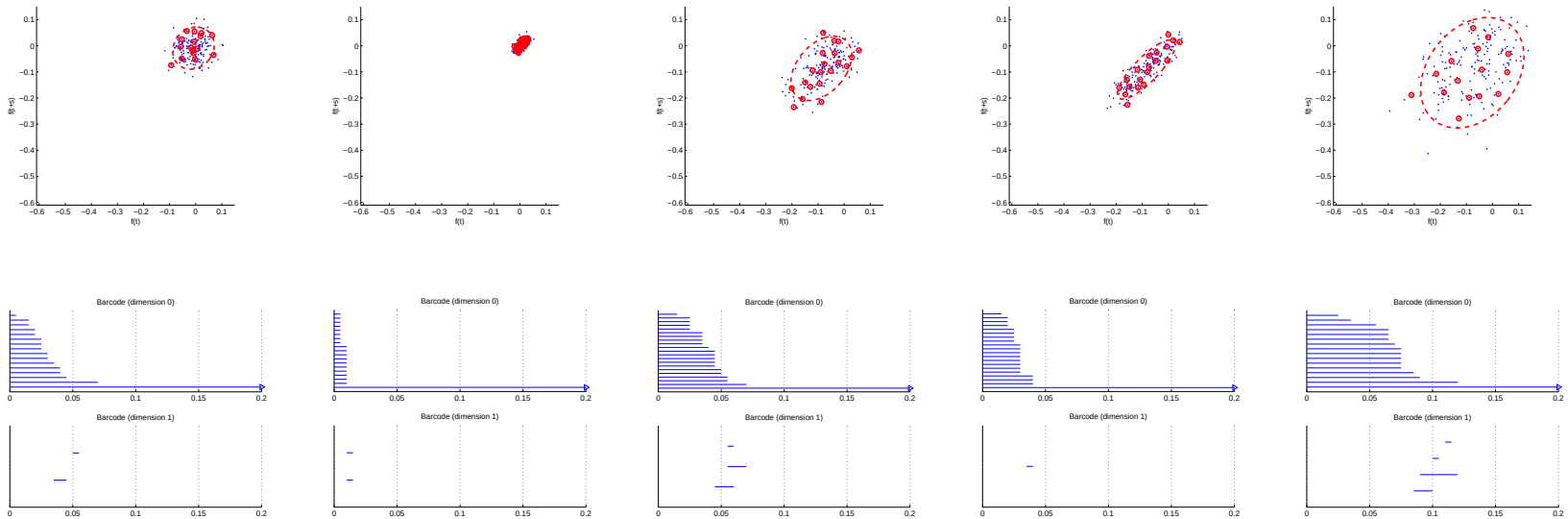


(e) Vietoris-Rips complex (f) Witness Rips complex

Figure: Comparison of Vietoris-Rips and Witness Rips complex with the same landmarks and r. Blue dots are synthetic point cloud and red circles are landmarks. (a) constructs the complex regardless those blue dots, so it cannot exactly reflect the structure of the point cloud; (b) utilizes the rest blue dots to estimate whether a simplex is reasonable to be there, thus the simplex without enough witnesses will be ignored.

Persistent Homology

Barcode and Betti



Which Classifier?

Try:

SVM, Gaussian process, kNN and Decision tree.

Choose:

Decision tree

Reason:

Less manual parameters, faster, suitable for small data set with high uncertainty.

Modified Decision Tree (MDT)

Feature selection at a node

Correlation between each feature and the ground truth is calculated, and the feature with the smallest P-value is selected to split the data.

Constrains of splitting

Our case is multi-class classification, so the data on two sides of the splitting boundary should depart from each other, and the splitting entropy should keep the lowest.

Objective Function

At certain node, the selected feature is denoted as f, and assume there are $n \geq 2$ categories. Mean of each class is $\boldsymbol{\mu} \in \mathbb{R}^n$ sorted by ascending order, and the corresponding standard deviation is stored in $\boldsymbol{\sigma} \in \mathbb{R}^n$. Thus the objective function of choosing the splitting boundary can be expressed as

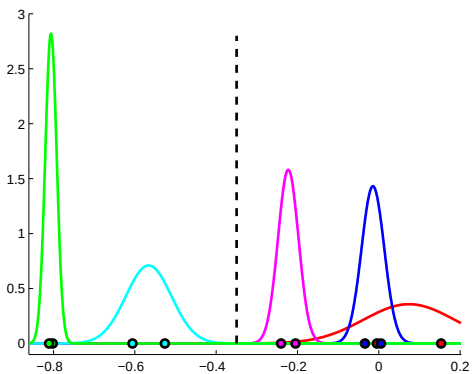
$$\begin{aligned} \arg\min_{x_i} & \left\{ \frac{G_{\mu_i,\sigma_i}(x_i)}{\mu_{i+1} - \mu_i} + \gamma E(\mathbf{f}, x_i) \right\} \\ \text{s.t.} \quad & G_{\mu_i,\sigma_i}(x_i) = G_{\mu_{i+1},\sigma_{i+1}}(x_i) \\ & \mu_i < x_i < \mu_{i+1}, \ i = 1, 2, \cdots, n-1 \end{aligned}$$

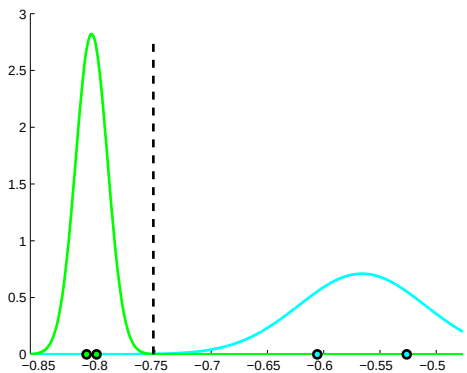
where $G_{\mu_i,\sigma_i}(\cdot)$ denotes the Gaussian function whose mean and variance are μ_i and σ_i^2 respectively.

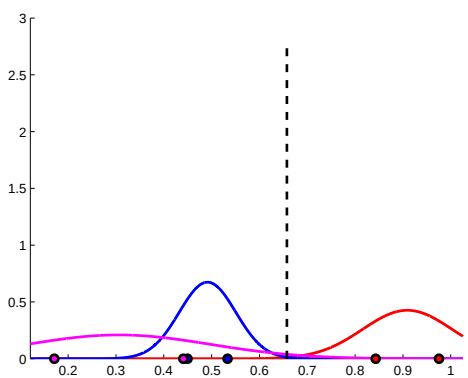
$$E(\mathbf{f}, x_i) = \begin{cases} -\sum_{c=1}^{n} \sum_{l=1}^{2} N_l^c \ln \frac{N_l^c}{N^c} & , \text{ no empty leaf} \\ 0 & , \text{ otherwise} \end{cases}$$

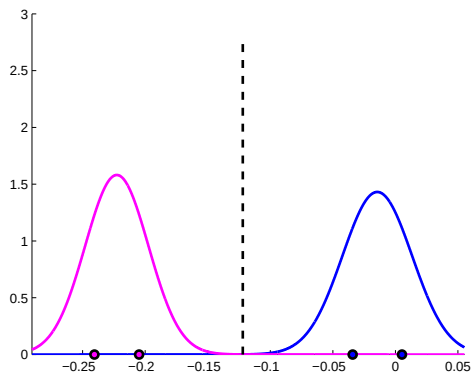
Construction of Decision Tree

A sample of constructing decision tree implemented. Left: root node. Middle: leaf nodes. Right: leaf node. Those pure nodes are not shown here.









Experiment Result

Leave-one-out cross validation

10 samples for each voltage level.

2 samples for each concentration level.

Voltage (mV)	20	40	70	135
MDT	<70%	80%	90%	100%
DT	—	—	—	30%
RF	—	—	—	50%
SVM	—	—	—	60%
GP	—	—	—	50%
kNN	—	—	—	30%
Unmixing	—	—	—	20%

Small data size makes the experiment result not so reliable.

Thank you!