Statistical Machine Learning Approaches to Change Detection

Song Liu, JSPS DC2,

Sugiyama Lab, Tokyo Institute of Technology

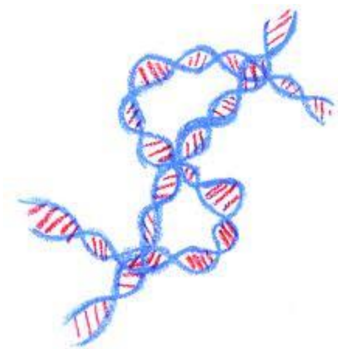
2014/3/10

Song Liu (Tokyo Tech)

- Graduating PhD student (3rd year) from Tokyo Tech
  - Thesis: Statistical Machine Learning Approaches to Change Detection
  - Fellow of Japan Society for Promotion of Science (JSPS)
- Visiting at National Institute of Informatics, JP
- MSc in Advanced Computing (University of Bristol)
- BEng in Computer Science (Soochow University, CN)

Learning from Data

- The era of big data is coming.



1 hour video per second

1M transaction per hour

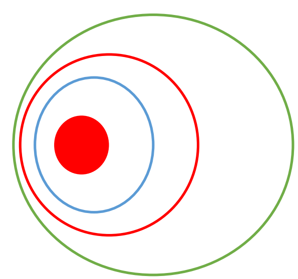
2.5 petabytes per 1000 person

- Without interpretation, raw data make no sense to human.
- Machine learning helps make sense out of data.

The Human Learning in Astronomy
A brief history

- Learning Constellations
  - Patterns of Stars
  - Predicting warfare, harvest

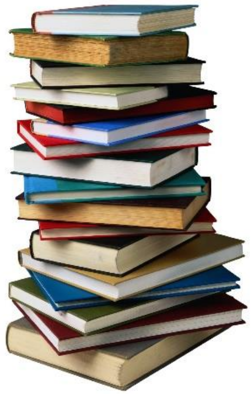
- Patterns of Planetary Movements
  - Tycho Brahe's data
  - Kepler's law

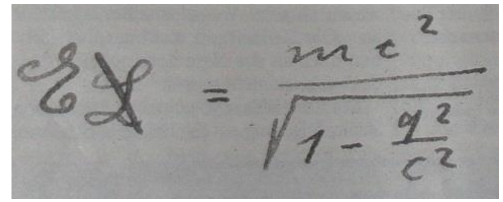


- Newton Laws
  - Predicting the movements of new planets

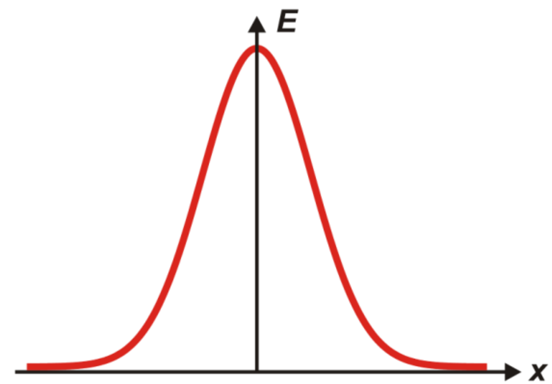
$$F = ma$$

Statistical Machine Learning





Data is so big, so we look for compressive rules (**patterns**).



Data involving uncertainty, we prefer **statistical** patterns .

However, Data is Always Changing!

- Patterns learned today may be useless tomorrow.

Everyday, 20% Google queries have never been seen before[1].

- **Learning pattern changes** are also important.

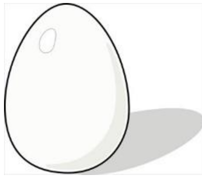
Red spots are detected changes on NASA satellite images, reflecting the heaviest damages after super typhoon Haiyan struck Philippines on Nov. 8 2013 [2].

- We need a new paradigm of machine learning.

1. http://certifiedknowledge.org/blog/are-search-queries-becoming-even-more-unique-statistics-from-google
2. http://www.nasa.gov/content/goddard/haiyan-northwestern-pacific-ocean/#.Uq3u6vQW2So

Two Paradigms of Machine Learning

(p.5)



Static Learning: Data do not change

Supervised

Predict Future: $p(y|\boldsymbol{x})$

Classification, Regression

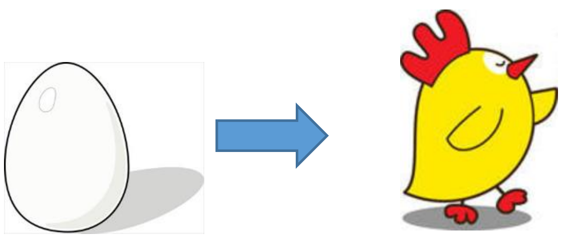
(Duda et al., 2001, Bishop, 2006)

Unsupervised

Learn Patterns: $p(\boldsymbol{x})$

Clustering, Anomaly Detection

(Murphy, 2012, Shaw-Taylor & Cristianini, 2004)



Dynamic Learning: Data are shifting

Supervised

Predict future under data shifts

$$p(y'|\boldsymbol{x}'), p \neq p'$$

e.g. Transfer learning (Sugiyama et al., 2012)

Unsupervised

Learn changes of patterns

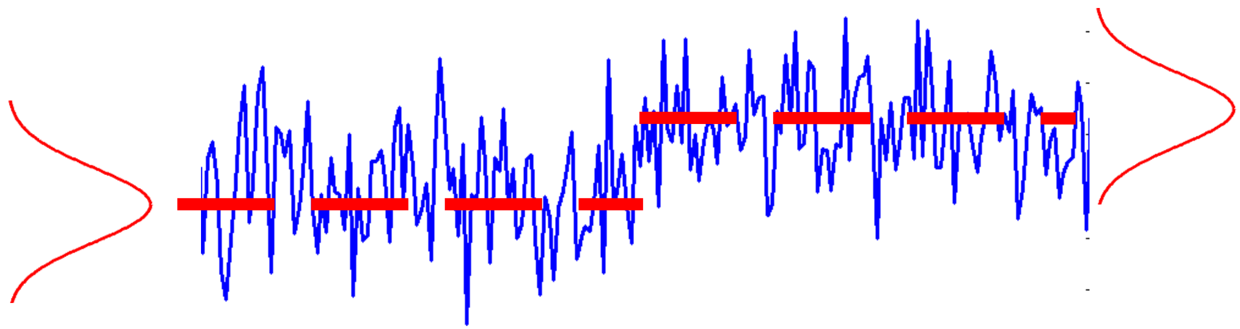
$$p/q \text{ or } p - q, p \neq q$$

Change detection

Two Ways of Learning Changes

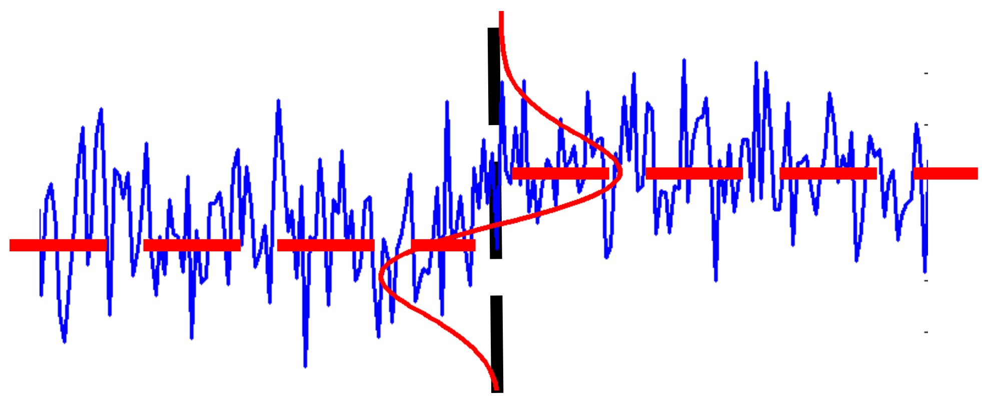
Separated Learning Approach

①



Direct Learning Approach

②

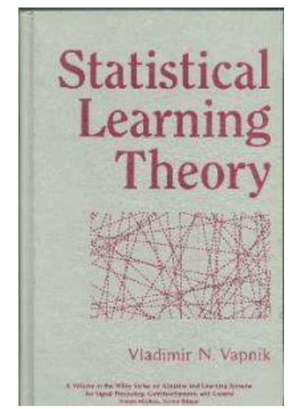


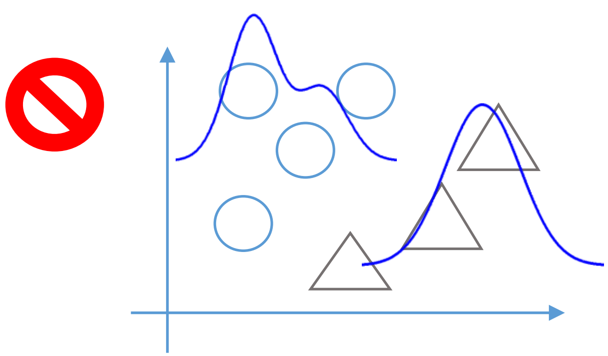
Which one is the best ?

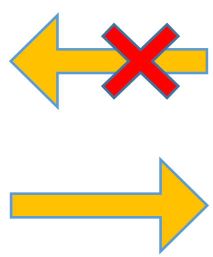
Vapnik’s Principle

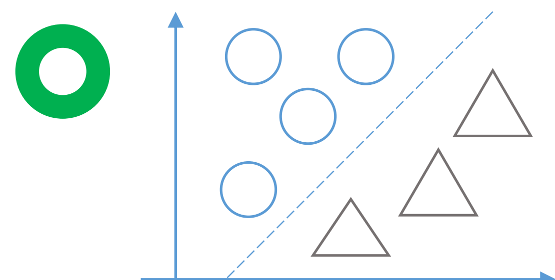
(Vapnik, Statistical Learning Theory, 1998)

- Generally, we prefer ② to ① because of **Vapnik’s principle**.
- “When solving a problem of interest, one should not solve a more general problem as an intermediate step.”







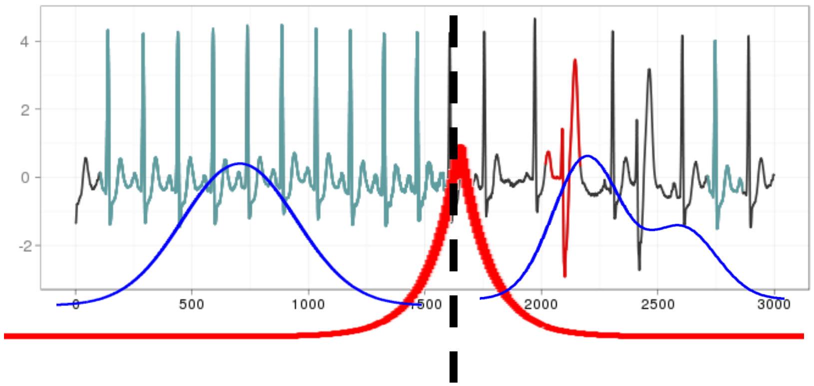


Support Vector Machines!

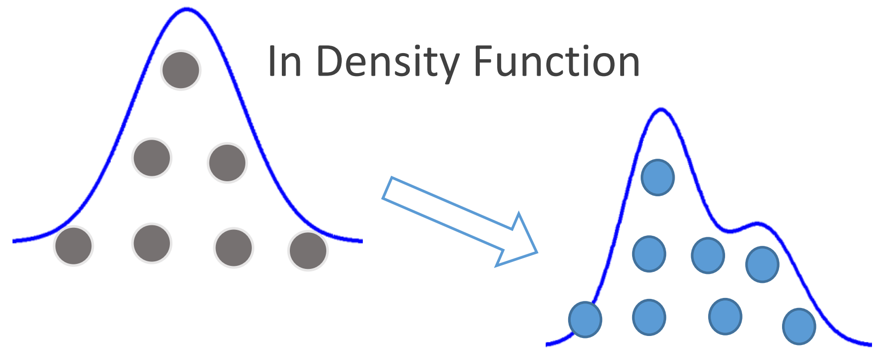
Unsupervised Change Detection

(p.10)

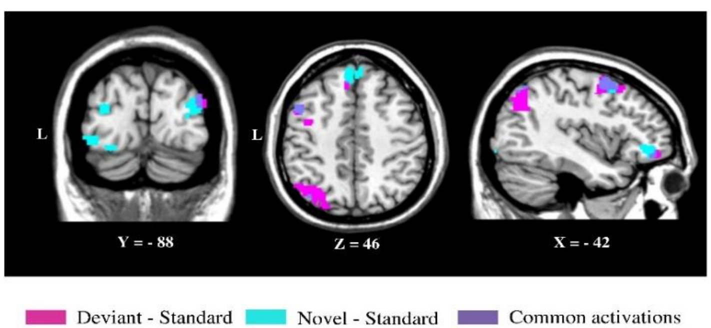
- Has it been changed? **Distributional Change Detection**



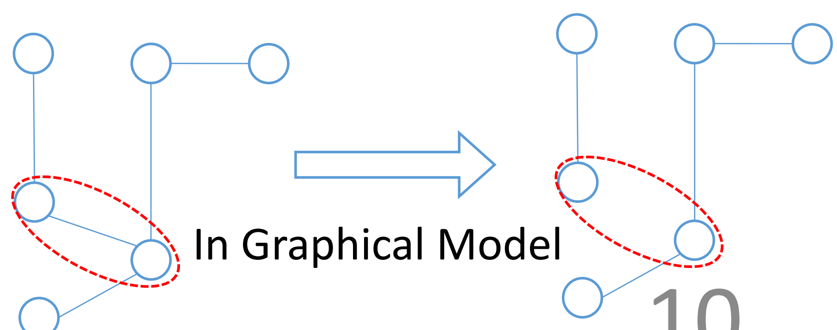
Heartbeat disorder detection, from JMOTIF project.



- What has been changed? **Structural Change Detection**



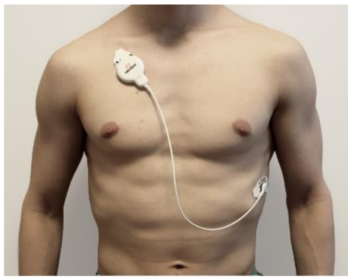
Changes in brain activation (Cléry et al., 2006)



Chapter 2,
Distributional Change Detection

1. Background & Existing Methods
2. Problem Formulation
3. Divergence based Change-point Score
4. Experiments

Changes in Time-series



Health monitoring

e.g. Kawahara et al., 2012

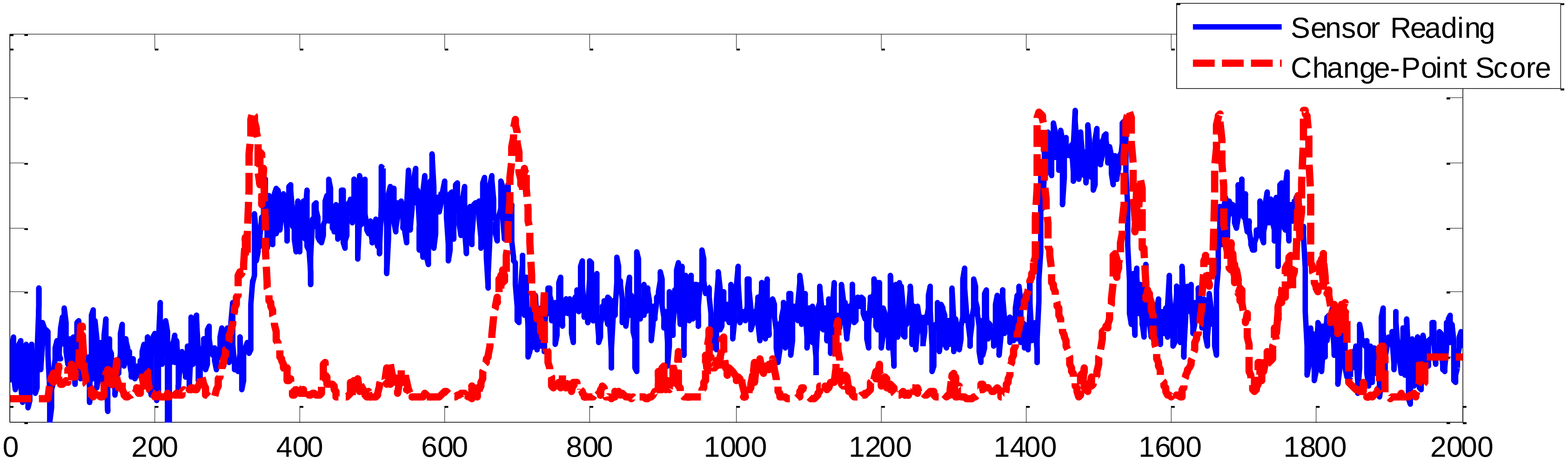
Engineering fault detection

e.g. Keogh et al., 2005

Network intrusion detection

e.g. Takeuchi et al., 2006

Distributional Change Detection in Time-series

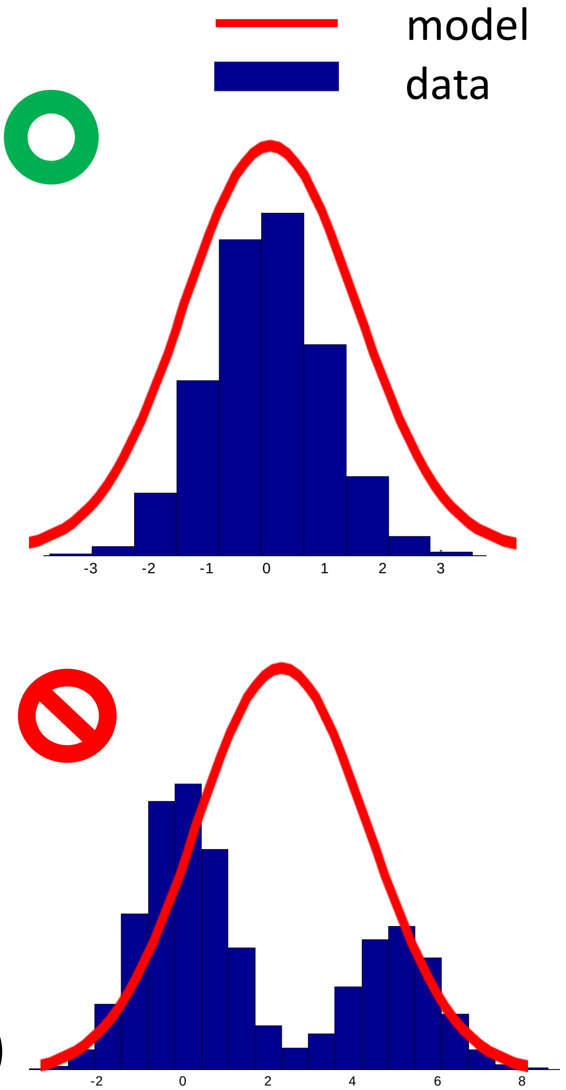


- **Objective**: Detecting abrupt changes lying among time-series data
- **Change-point score**: Plausibility of changes that have happened
- **Metrics**: True positive rate, false positive rate, and detection delay

Existing Methods

- Model based Methods:
  - **Auto Regressive (AR)**
    - (Takeuchi and Yamanishi, 2006)
  - Autoregressive model with Gaussian noise.
  - **Singular Spectrum Transformation (SST)**
    - (Moskvina and Zhigljavsky, 2003)
  - **Subspace Identification (SI)**
    - (Kawahara et al., 2007)
  - State Space model with Gaussian noise.

- Model-free Methods:
  - **One-class Support Vector Machine (OSVM)**
    - (Desobry et al., 2005)
  - **Density Ratio based Methods (KLIEP)**
    - KLIEP (Kawahara & Sugiyama, 2009)
  - **No model assumption.**



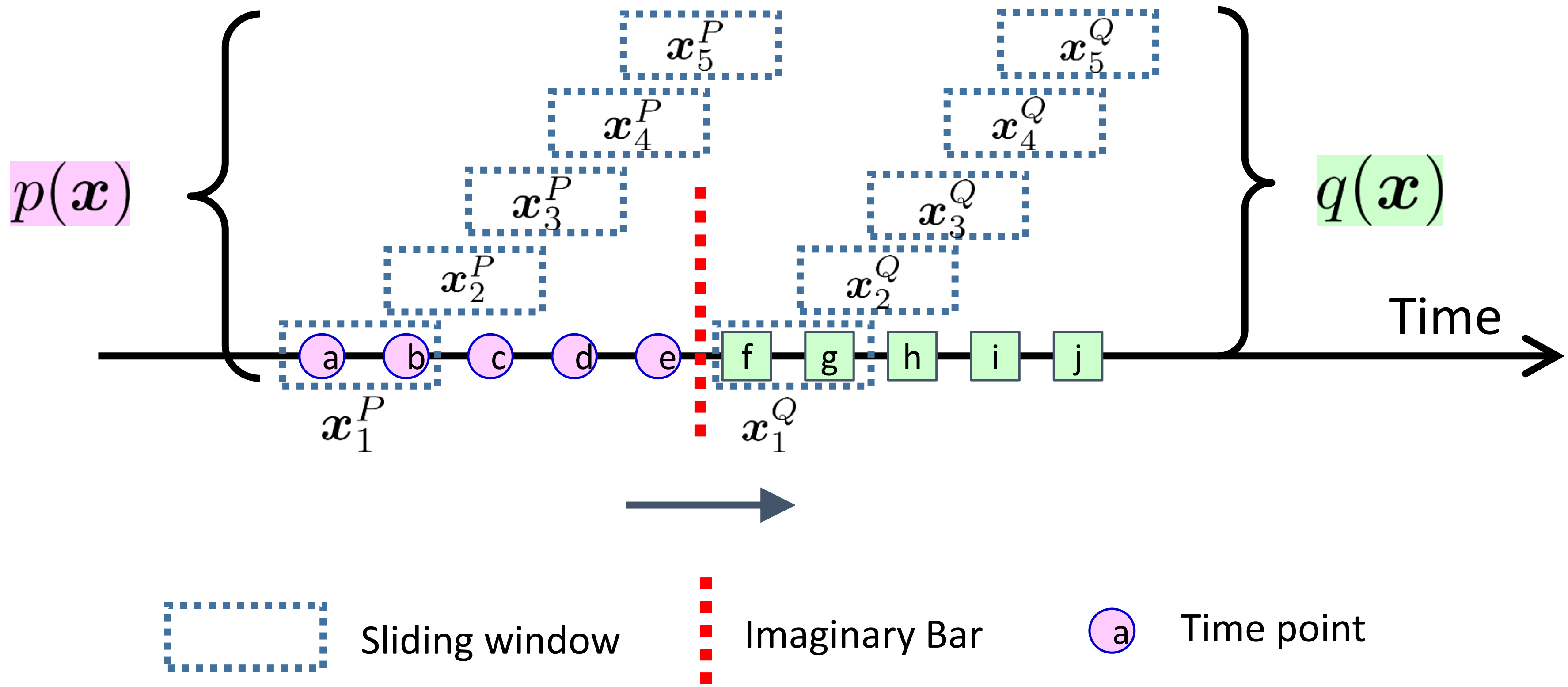
Model based methods fail to characterize data when model mismatches.

Chapter 2, Distributional Change Detection

1. Background & Existing Methods
2. Problem Formulation
3. Proposed Method
4. Experiments

Formulate Problem from Time-series

- Construct samples by using sliding window.
- An imaginary bar in the middle divides samples into two groups.
- Assume two groups of samples are from p and q.



How to generate a change point score?

Chapter 2,
Distributional Change Detection

1. Background & Existing Methods
2. Problem Formulation
3. Proposed Method
4. Experiments

Divergence based Change Point Detection

- Need a **measure of change** between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$.
- f-divergence is a **distance** from p to q.

(Ali and Silvey, 1966)

$$\mathrm{D}[p||q] = \int f\left(\boxed{\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}}\right) q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

f is convex

$f(1) = 0$

- f is a function of a ratio between two **probability density functions (**p **and** q**)**.
- $\mathrm{D}[p||q]$ is asymmetric, thus we symmetrized it:

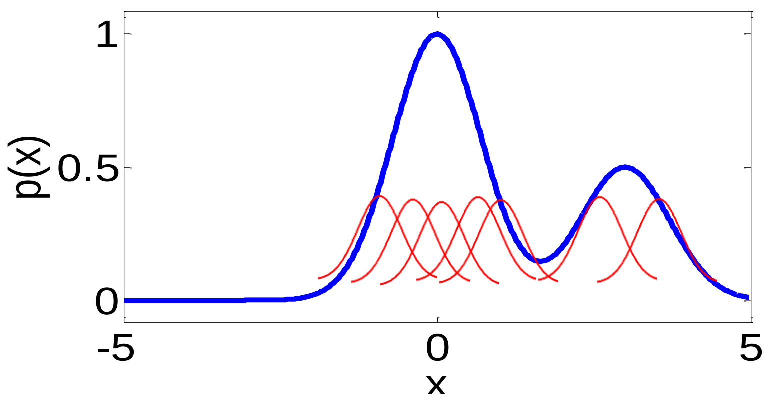
$$\mathrm{D}[p||q] + \mathrm{D}[q||p]$$

Density Estimation

- A naive way is a two step approach:

$$\left.\begin{array}{l} ① \; p(\boldsymbol{x}) \approx \hat{p}(\boldsymbol{x}) \\ ② \; q(\boldsymbol{x}) \approx \hat{q}(\boldsymbol{x}) \end{array}\right\} \quad \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \approx \frac{\hat{p}\,(\boldsymbol{x})}{\hat{q}\,(\boldsymbol{x})}$$

- e.g. Kernel Density Estimation (Csörgo & Horváth 1988)

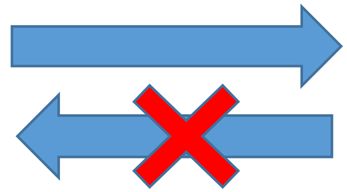


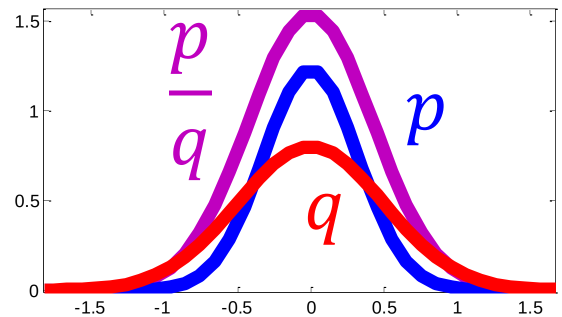
$$\hat{p}\,(\boldsymbol{x};\sigma) = \frac{1}{n}\sum_{i=1}^{n} K_\sigma(x, x_i)$$

$$K_\sigma(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{-2\sigma^2}\right)$$

- ① or ② is carried out without taking care of the ratio.
  - Error may be magnified after the division.

Density Ratio Estimation

Knowing p and q  Knowing $\frac{p}{q}$



Vapnik's Principle!

- ***Direct*** Density Ratio Estimation (Sugiyama et al., 2012).

$$\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \approx \hat{r}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i}^{n} \theta_i K_\sigma(\boldsymbol{x}, \boldsymbol{x}_i)$$

- By plugging in estimated density ratio, we may obtain various *f*-divergences.

Kullback-Leibler Divergence as Change-point Score (p.22)

We may have different choices for change-point score:

- Kullback-Leibler Divergence (Kullback and Leibler, 1951)

$$\mathrm{KL}[p||q] = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$

- Has been applied, and promising results were obtained.

Kawahara & Sugiyama, SDM 2009

- Density ratio can be obtained via **Kullback-leibler Importance Estimation Procedure (KLIEP)** (Sugiyama, et al., NIPS 2008).

○ Optimization for density ratio is convex!

🚫 Slow, less robust

Pearson Divergence as Change-point Score (pp.24)

- The log term in Kullback-leibler divergence can go infinity if $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} = 0$.
- Pearson Divergence (Pearson, 1900)

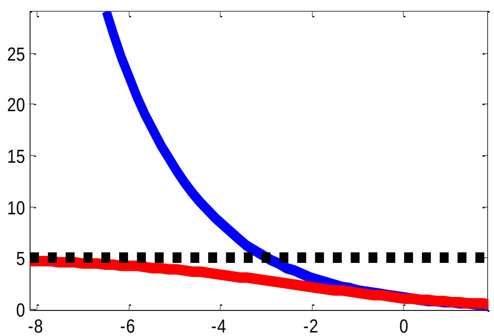
$$\mathrm{PE}[p||q] = \frac{1}{2}\int q(\boldsymbol{x})\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right)^2 \mathrm{d}\boldsymbol{x}$$

  - Density ratio can be obtained via **unconstrained Least Square Importance Fitting (uLSIF)** (Kanamori, et al., JMLR 2009).
  - Solution has an analytic form.

○ Fast, relatively robust!

Relative Pearson Divergence as Change-point Score (pp.26)

- The density ratio may go unbounded.
- Bounded Density Ratio -> Relative Density Ratio



$\frac{p}{q}$ $\qquad$ $\frac{p}{\alpha p+(1-\alpha)q} \leq \frac{1}{\alpha}$

- Relative Pearson Divergence (Yamada, et al. , NIPS 2011)

$$\mathrm{PE}_\alpha[p\|q] = \mathrm{PE}[p\|\alpha p + (1-\alpha)q]$$
$$= \frac{1}{2}\int (\alpha p(\boldsymbol{x}) + (1-\alpha)q(\boldsymbol{x}))\left(\frac{p(\boldsymbol{x})}{\alpha p(\boldsymbol{x}) + (1-\alpha)q(\boldsymbol{x})} - 1\right)^2 \mathrm{d}\boldsymbol{x}$$

More robust!

Divergences as Change-point Scores (Summary) (pp.22-28)

- **Kullback-Leibler Divergence (via KLIEP) :**

$$\widehat{\mathrm{KL}} := \frac{1}{n}\sum_{i=1}^{n}\log \hat{r}(\boldsymbol{x}_i)$$

Existing method

Previous Work: Kawahara & Sugiyama, 2009

- **Pearson Divergence (via uLSIF):**

$$\widehat{\mathrm{PE}} := -\frac{1}{2n}\sum_{i=1}^{n}\hat{r}(\boldsymbol{x}'_i)^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{r}(\boldsymbol{x}_i) - \frac{1}{2}$$

- **Relative Pearson Divergence (via RuLSIF):**

$$\widehat{\mathrm{PE}}_\alpha = -\frac{\alpha}{2n}\sum_{i=1}^{n}\hat{r}(\boldsymbol{x}_i)^2 - \frac{1-\alpha}{2n}\sum_{i=1}^{n}\hat{r}(\boldsymbol{x}'_i)^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{r}(\boldsymbol{x}_i) - \frac{1}{2}$$

Robust Change Detection

Liu et al., SPR 2012, Neural Networks 2013

Chapter 2, Distributional Change Detection

1. Background & Existing Methods
2. Problem Formulation
3. Proposed Method
4. Experiments

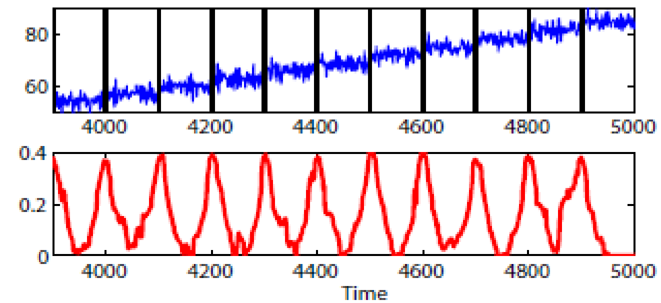
Experiments (Synthetic)

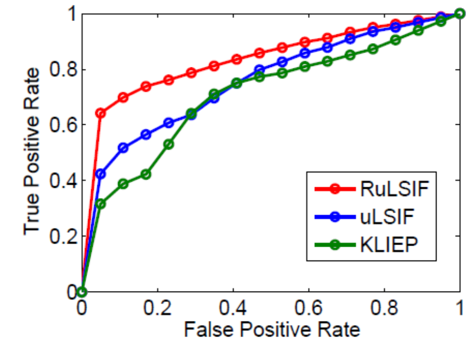
original signal

score

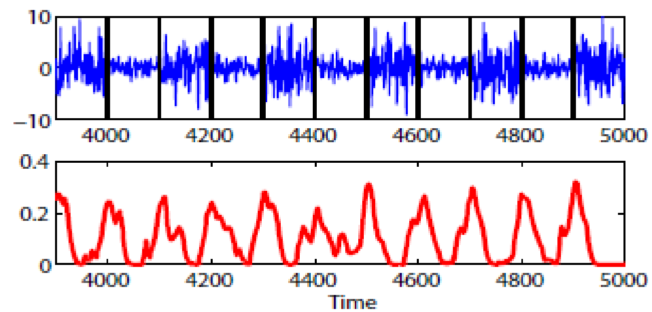
ground truth

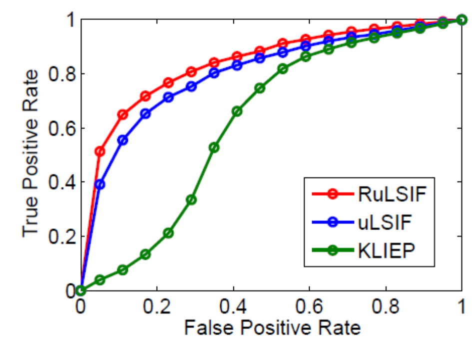
Mean Shift



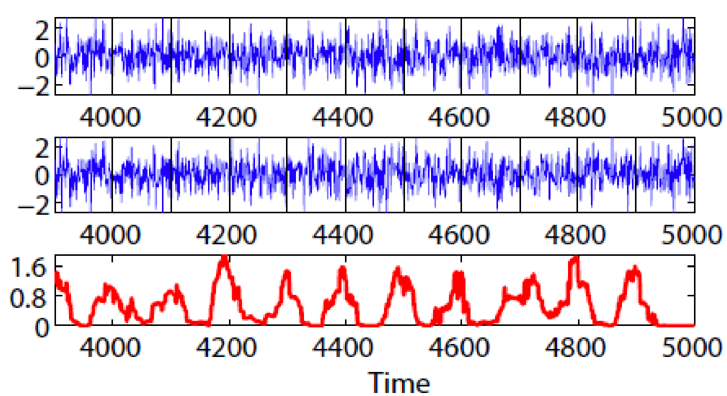


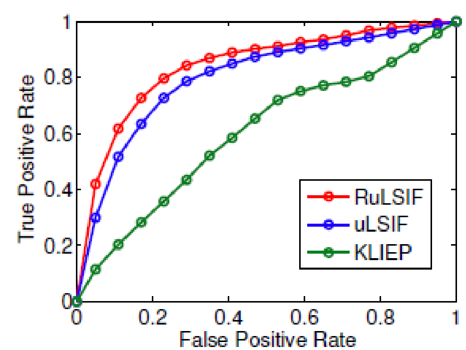
Variance Scaling



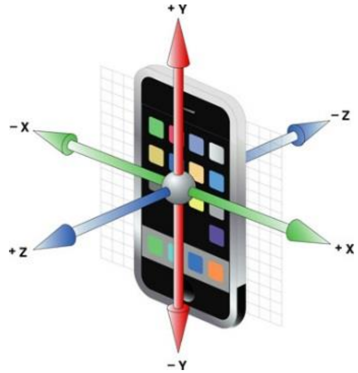


Covariance Switching

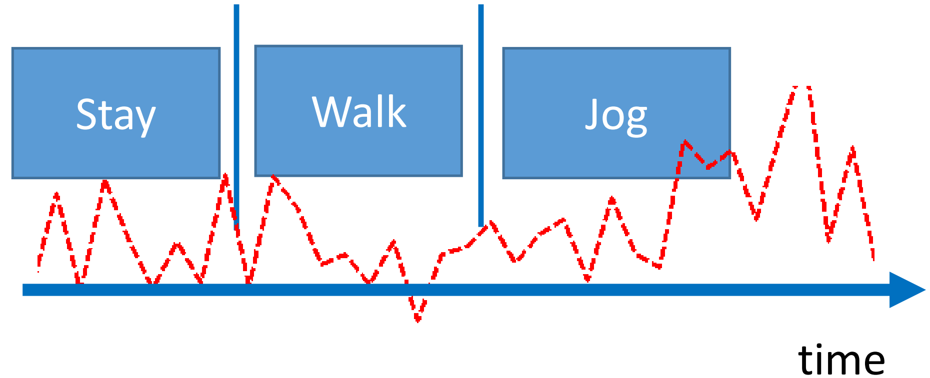




Experiments (Sensor)



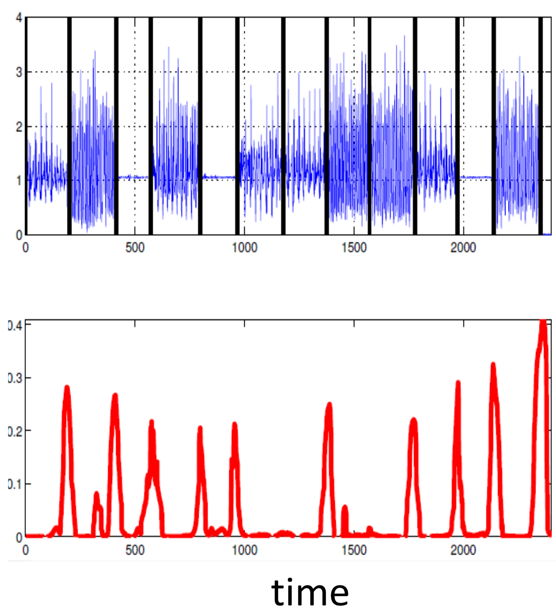
HASC 2011 Challenge

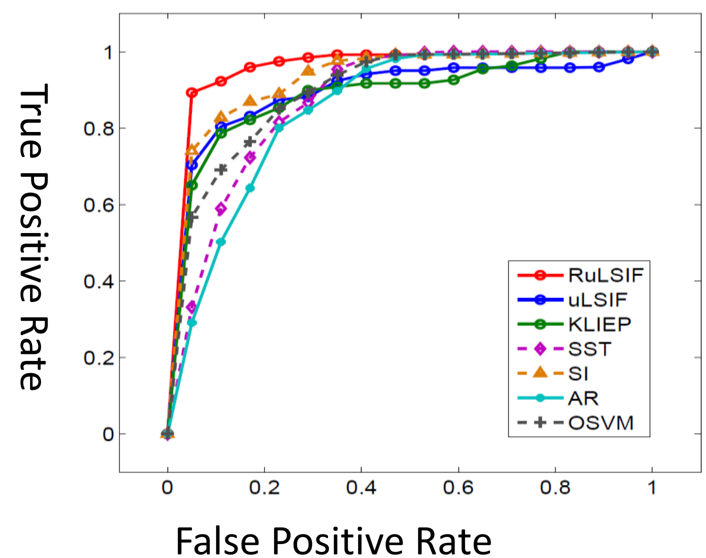


Mobile *Sensor* data, measured in 20 seconds, recording human activity.

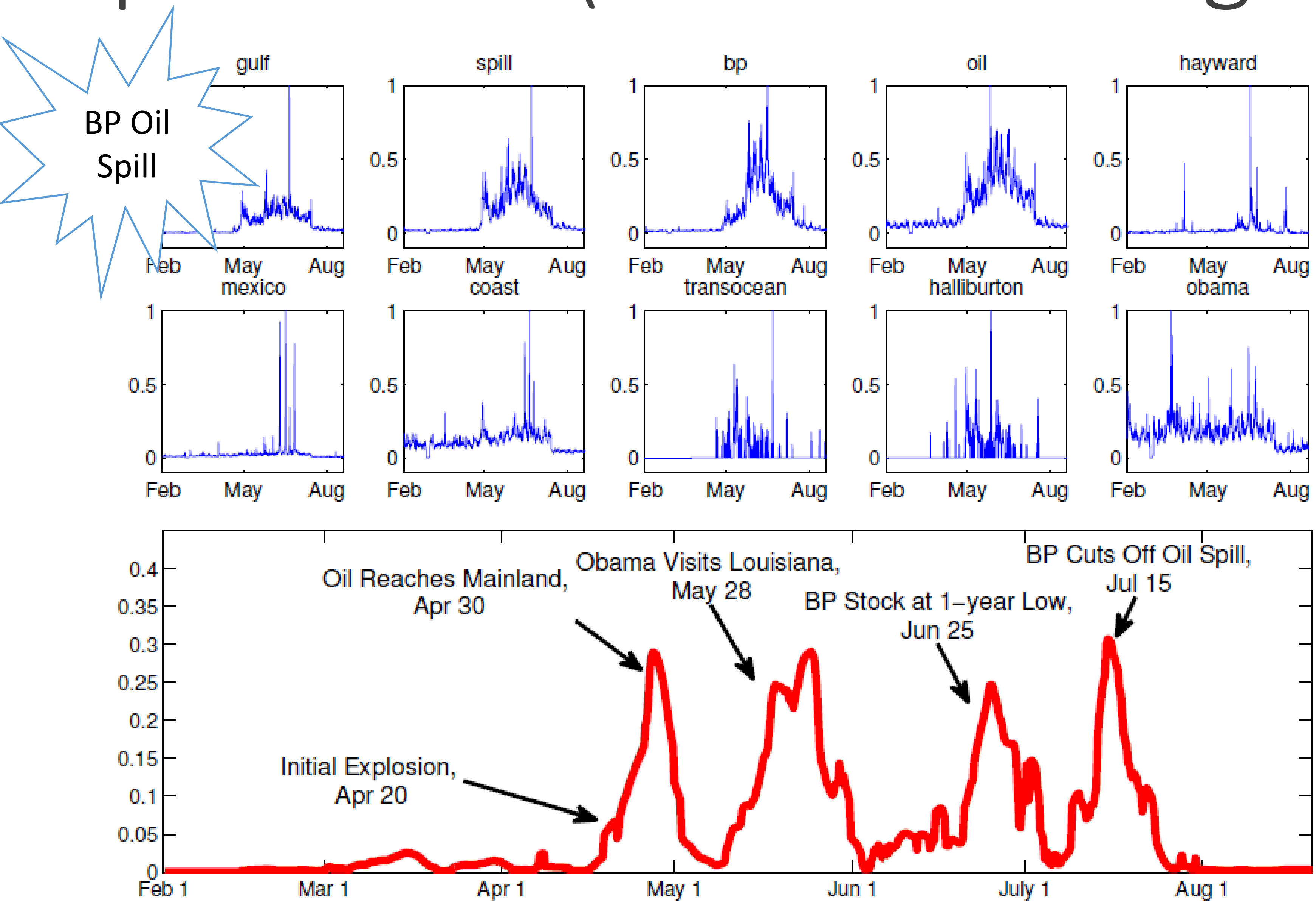
Task: Segmenting the activities.

e.g. walk, jogging, stairs up/down…





Experiments (Twitter Messages)



Conclusion

- A **robust** and **model-free** method for distributional change detection in time-series data was proposed.

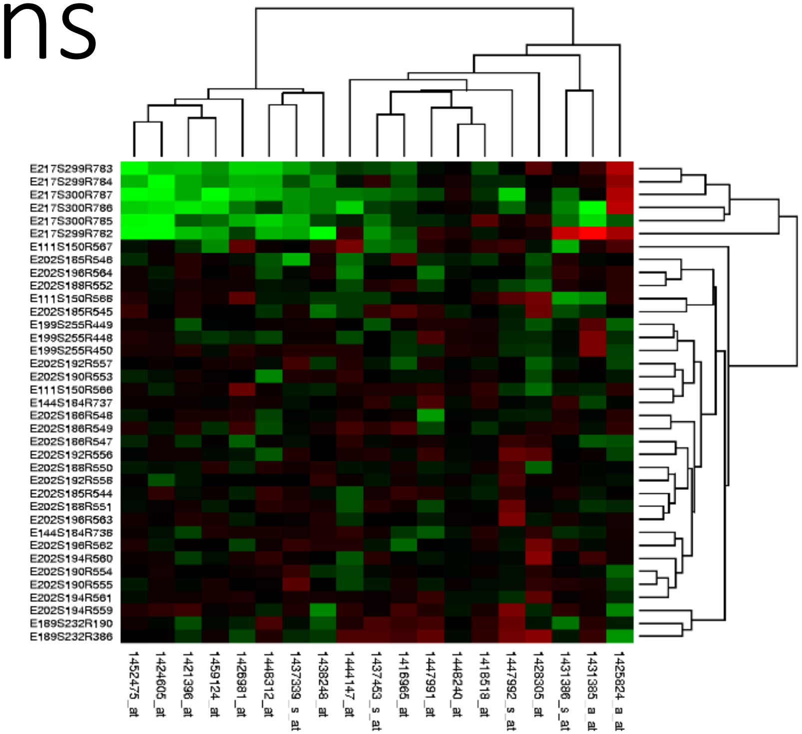
- The proposed method obtained the **leading** performance in various artificial and real-world experiments.

Chapter 3,
Structural Change Detection

1. Background
2. Separate Estimation Methods
3. Proposed Methods
4. Experiments

Patterns of Interactions

- Genes regulate each other via gene network.



(Wikipedia)

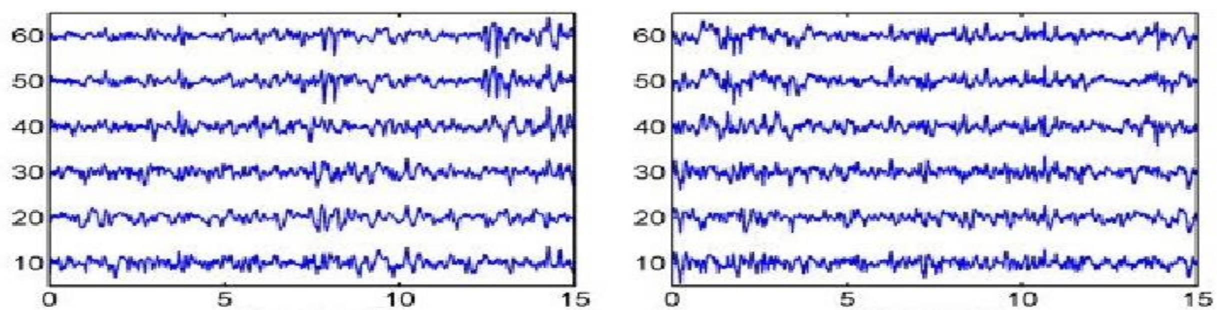
- Brain EEG signals may be synchronized in a certain pattern.

- However, such patterns of interactions may be **changing!**

Structural Change Detection

- Interactions between features may change. $\boldsymbol{x} = \left(x^{(1)}, \dots, x^{(d)}\right)^{\mathrm{T}}$

$$\{\boldsymbol{x}_i^P\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) \qquad \{\boldsymbol{x}_i^Q\}_{i-1}^n \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$$



(Williamson et al., 2012)

The change of brain signal correlation at two different experiment intervals.

- e.g. Interactions between some genes may be activated,
  - but only under some conditions.
- “apple” may co-occur with “banana” quite often in cookbook,
  - but **not in IT news**.

Chapter 3,
Structural Change Detection

1. Background
2. Existing Methods
3. Proposed Methods
4. Experiments

Gaussian Markov Networks (GMNs)

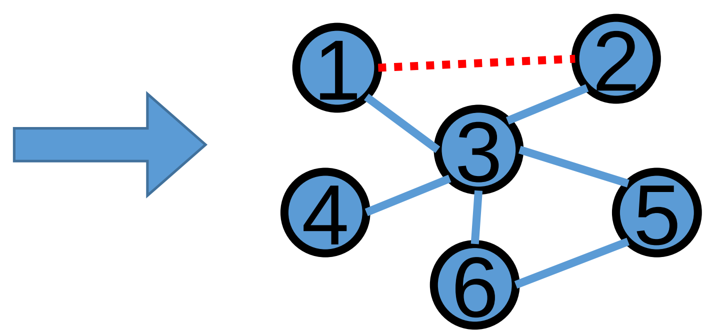
- The interactions between random variables can be modelled by **Markov Networks.**
- **Markov Networks (MNs)** are **undirected** Graphical Models.
- The simplest example of MN is a Gaussian MN:

$$p(\boldsymbol{x}; \boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{x}\right)$$

$\boldsymbol{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$ $\boldsymbol{\Theta}$ is the inverse covariance matrix

- We can visualize the above MN using an undirected graph.

$\boldsymbol{\Theta}_{1,2} = 0$



Estimating Separate GMN

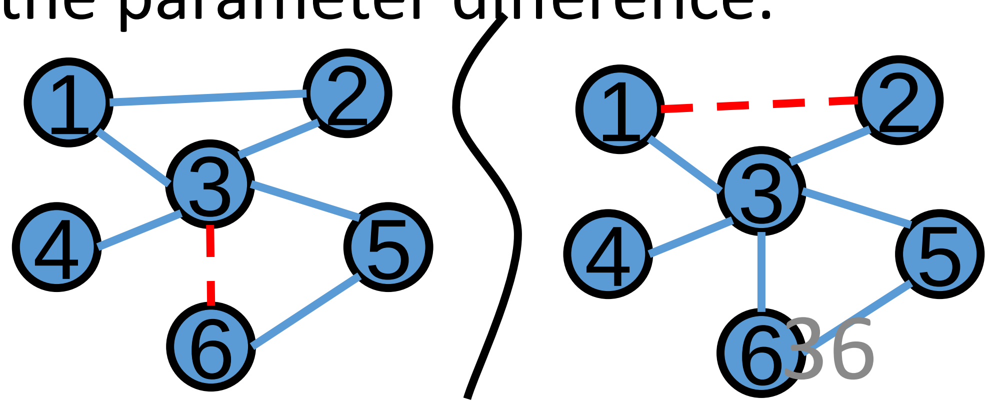
Tibshirani, JRSS 1996; Friedman et al., Biostatistics 2008

- Recall, we would like to detect changes between MNs.
- Changes can be found once the structure of two **separate** GMNs are known to us.
- Estimating Sparse GMNs can be done via **Graphical Lasso (Glasso)**.

$$\max_{\boldsymbol{\Theta}^P} \sum_{i=1}^{n} \log p(\boldsymbol{x}_i^P; \boldsymbol{\Theta}^P) - \lambda^P \|\boldsymbol{\Theta}^P\|_1$$

- Idea: Twice Glasso, take the parameter difference.

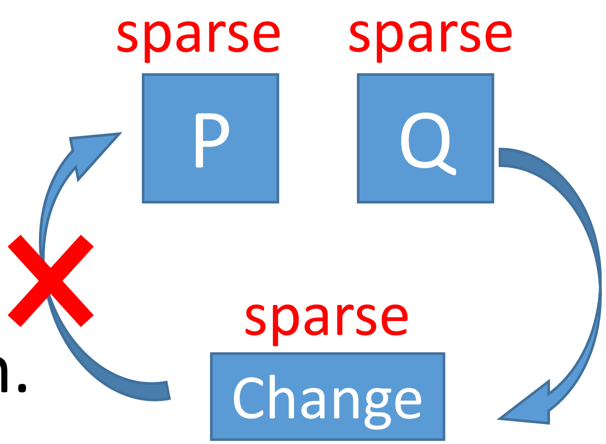
$$\boxed{\boldsymbol{\Theta}_{u,v}^P - \boldsymbol{\Theta}_{u,v}^Q}$$



Glasso: Pros and Cons

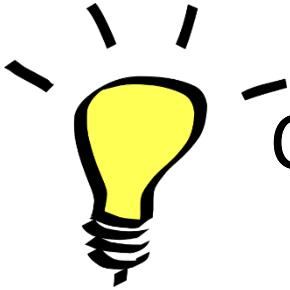
☺Pros:

- Statistical properties are well studied.
- Off-the-shelf software can be used.
- **Sparse change is produced.**



☹ Cons:

- Cannot detect high-order correlation.
- Not clear how to choose hyper-parameters.
- **Does not work if p or q is dense.**



Can we combine two optimizations into one?

Semi-direct Approach (Fused-lasso)

- We can impose sparsity directly on $\boldsymbol{\Theta}^{P}_{u,v} - \boldsymbol{\Theta}^{Q}_{u,v}$, using **Fused-lasso (Flasso)** (Tibshirani et al., 2005).

- Consider the following objective:

$$\max_{\boldsymbol{\theta}} \underbrace{\ell(\boldsymbol{x}^P, \boldsymbol{\Theta}^P)}_{\text{Gaussian}} + \underbrace{\ell(\boldsymbol{x}^Q, \boldsymbol{\Theta}^Q)}_{\text{Gaussian}} - \boxed{\lambda \|\boldsymbol{\Theta}^P - \boldsymbol{\Theta}^Q\|_1}$$

(Zhang & Wang, **UAI2010**)

☺ We don't have to assume p or q is sparse.

☺ Sparsity control is much easier than Glasso.

☹ Gaussianity is still assumed.

Nonparanormal (NPN) Extension

Liu et al., JMLR 2009

- Gaussian assumption is too restrictive.
- However, considering general non-Gaussian model can be computationally challenging.
- We may consider a method half-way between Gaussian and non-Gaussian.
- Non-paranormal (NPN).

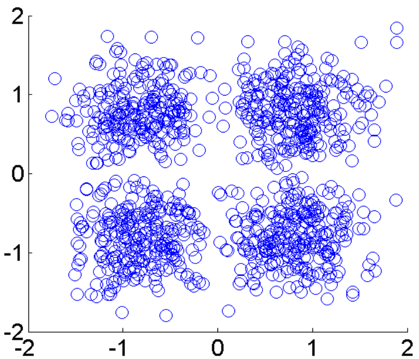
$$p(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\Theta} \boldsymbol{f}(\boldsymbol{x})\right) \prod_{i=1}^{d} |f_i'(x^{(i)})|$$

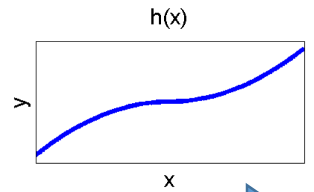
$$\boldsymbol{x} = (x^{(1)}, x^{(2)}, \dots, x^{(d)})$$

$$\boldsymbol{f}(\boldsymbol{x}) = (f_1(x^{(1)}), f_2(x^{(2)}), \dots, f_d(x^{(d)}))$$

f_k : Monotone, differentiable function

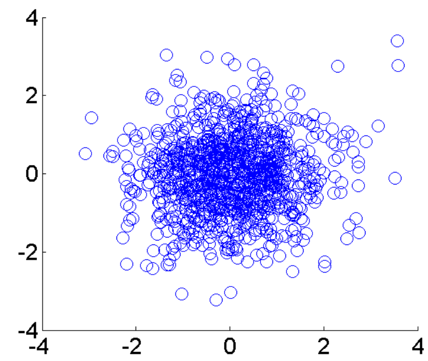
Nonparanormal (NPN) Extension





$$f(x) = \text{sign}(x)x^2$$

$$f^{-1}(x) = \text{sign}(x)\sqrt{|x|}$$



☺ More flexible than Gaussian methods, still tractable.

☹ However, NPN extension is still restrictive.

Non-Gaussian Log-linear Model

- Pairwise Markov Network

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{f}(x^{(u)}, x^{(v)})\right)$$

- $\boldsymbol{f}$ are feature vectors.

$\boldsymbol{f}: \mathcal{R}^2 \to \mathcal{R}^b$

Gaussian: $f_{gau}(x, y) = xy$

Nonparanormal: $f_{npn}(x, y) = f(x)f(y)$

Polynomial: $\boldsymbol{f}_{poly}(x, y) = [x^k, y^k, x^{k-1}y \; \dots, x, y, 1]$

- The normalization term $Z(\boldsymbol{\theta})$ is generally intractable.
  - Gaussian or Nonparanormal models are exceptions.

The Normalization Issue

$$Z(\boldsymbol{\theta}) = \int \exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{f}(x^{(u)}, x^{(v)})\right) \mathrm{d}\boldsymbol{x}$$ **!**

- However, for a generalized Markov Network, there is no closed-form for $Z(\boldsymbol{\theta})$.
- Importance Sampling (IS):

$$Z(\boldsymbol{\theta}) = \int p_{\mathrm{inst}}(\boldsymbol{x}) \frac{\exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{f}(x^{(u)}, x^{(v)})\right)}{p_{\mathrm{inst}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$

- IS can be used to extend both Glasso and Flasso.
  - IS-Glasso, IS-Flasso.

Chapter 3,
Structural Change Detection

1. Background
2. Existing Methods
3. Proposed Methods
4. Experiments

Modeling Changes Directly

- Recall, our interest is: $\boldsymbol{\beta}_{u,v} := \boldsymbol{\theta}^P_{u,v} - \boldsymbol{\theta}^Q_{u,v}$

- The ratio of two MNs naturally incorporates the $\boldsymbol{\theta}^P_{u,v} - \boldsymbol{\theta}^Q_{u,v}$!

$$\frac{p(\boldsymbol{x};\boldsymbol{\theta}^P)}{q(\boldsymbol{x};\boldsymbol{\theta}^Q)} \propto \exp\left(\sum_{u \geq v} (\boldsymbol{\theta}^P_{u,v} - \boldsymbol{\theta}^Q_{u,v})^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right)$$

$$\exp\left(\sum_{u \geq v} \boldsymbol{\beta}^\top_{u,v} \boldsymbol{f}(x^{(u)}, x^{(v)})\right)$$

So, model the **ratio** directly!

Modeling Changes Directly

- Density ratio model:

$$r(\boldsymbol{x};\boldsymbol{\beta}) = \frac{1}{N(\boldsymbol{\beta})}\exp\left(\sum_{u\geq v}\boldsymbol{\beta}_{u,v}^{\top}\boldsymbol{f}(x^{(u)},x^{(v)})\right)$$

- The normalization term is:

$$N(\boldsymbol{\beta}) = \int q(\boldsymbol{x})\exp\left(\sum_{u\geq v}\boldsymbol{\beta}_{u,v}^{\top}\boxed{\boldsymbol{f}(x^{(u)},x^{(v)})}\right)\,\mathrm{d}\boldsymbol{x}$$

To ensure:

$$\hat{p} = q(\boldsymbol{x})r(\boldsymbol{x},\widehat{\boldsymbol{\beta}})$$

$$\int r(\boldsymbol{x};\boldsymbol{\beta})q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$$

Sample average approximation

Also works when **integral has no closed form**!

$$\frac{1}{n}\sum_{i}^{n}\exp\left(\sum_{u\geq v}\boldsymbol{\beta}_{u,v}^{\top}\boldsymbol{f}(x^{(u)Q},x^{(v)Q})\right)$$

Estimating Density Ratio

Sugiyama et al., NIPS 2007

- Kullback-Leibler Importance Estimation Procedure (KLIEP):

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{KL}[p\|\hat{p}] &= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\hat{p}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})\hat{r}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= \mathrm{Const.} - \boxed{\int p(\boldsymbol{x}) \log \hat{r}(\boldsymbol{x};\boldsymbol{\beta}) \mathrm{d}\boldsymbol{x}}\end{aligned}$$

$$\hat{p}(\boldsymbol{x}) = q(\boldsymbol{x})\hat{r}(\boldsymbol{x})$$

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \ell_{\mathrm{KLIEP}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{r}(\boldsymbol{x}_i;\boldsymbol{\beta})$$

Unconstrained convex optimization!

Tsuboi et al, JIP 2009

Sparse Regularization

L2 regularizers Group lasso regularizer

$$\max_{\boldsymbol{\beta}} \left[\ell_{\mathrm{KLIEP}}(\boldsymbol{\beta}) - \lambda_1 \boxed{\|\boldsymbol{\beta}\|_2^2} - \lambda_2 \boxed{\sum_{u \geq v} \|\boldsymbol{\beta}_{u,v}\|_2} \right]$$

Elastic Net

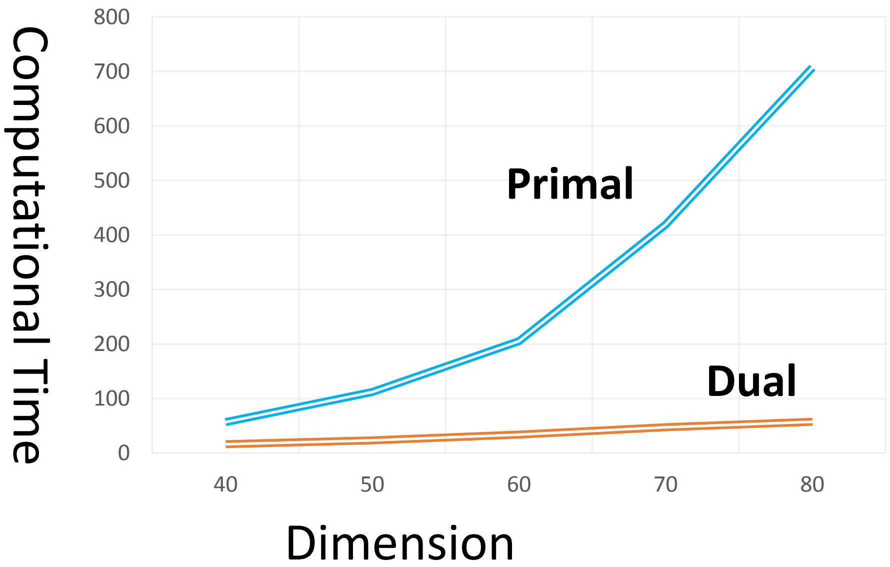
- Impose sparsity constraints on each factor $\boldsymbol{\beta}_{u,v}$.
  - equals to impose sparsity on changes.
- So finally, we can obtain a $\boldsymbol{\beta}$ with group sparsity!

The Dual Formulation p. 57

- When dimensionality is high, the dual formulation is preferred.

$$\min_{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{n_Q})^\top} \sum_{i=1}^{n_Q} \alpha_i \log \alpha_i + \frac{1}{\lambda_1} \sum_{u \geq v} \max(0, \|\boldsymbol{\xi}_{u,v}\| - \lambda_2)^2$$

$$\text{subject to } \alpha_1, \ldots, \alpha_{n_Q} \geq 0 \text{ and } \sum_{i=1}^{n_Q} \alpha_i = 1,$$



$$\boldsymbol{\theta}_{u,v} = \begin{cases} \frac{1}{\lambda_1}\left(1 - \frac{\lambda_2}{\|\boldsymbol{\xi}_{u,v}\|}\right)\boldsymbol{\xi}_{u,v} & \text{if } \|\boldsymbol{\xi}_{u,v}\| > \lambda_2, \\ \mathbf{0} & \text{if } \|\boldsymbol{\xi}_{u,v}\| \leq \lambda_2. \end{cases}$$

$$\boldsymbol{\xi}_{u,v} = \boldsymbol{g}_{u,v} - \boldsymbol{H}_{u,v}\boldsymbol{\alpha},$$

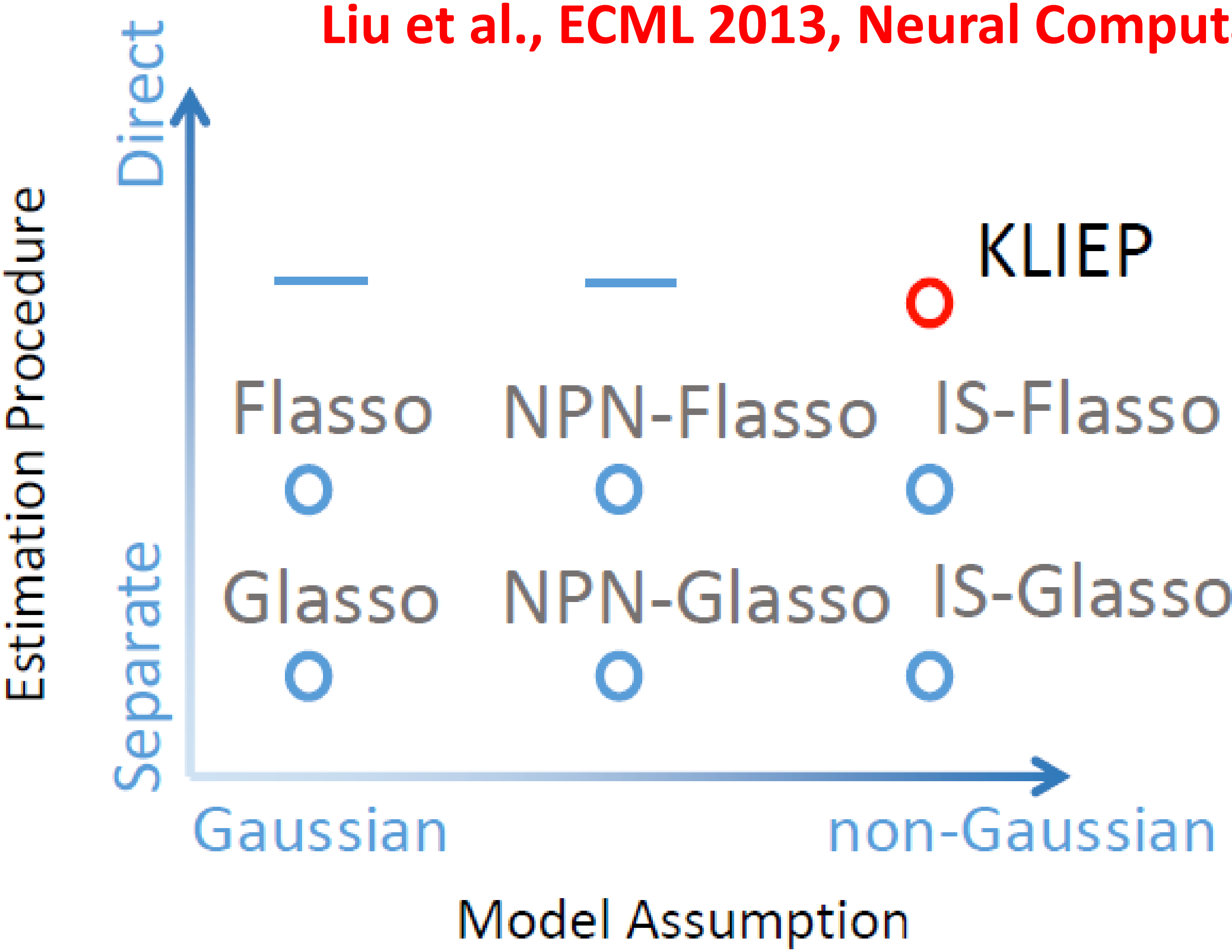
$$\boldsymbol{H}_{u,v} = [\boldsymbol{f}(x_1^{(u)Q}, x_1^{(v)Q}), \ldots, \boldsymbol{f}(x_{n_Q}^{(u)Q}, x_{n_Q}^{(v)Q})],$$

$$\boldsymbol{g}_{u,v} = \frac{1}{n_P}\sum_{i=1}^{n_P} \boldsymbol{f}(x_i^{(u)P}, x_i^{(v)P}).$$

We can recover primal solution!

Summary: Comparison between Methods

Liu et al., ECML 2013, Neural Computation 2014

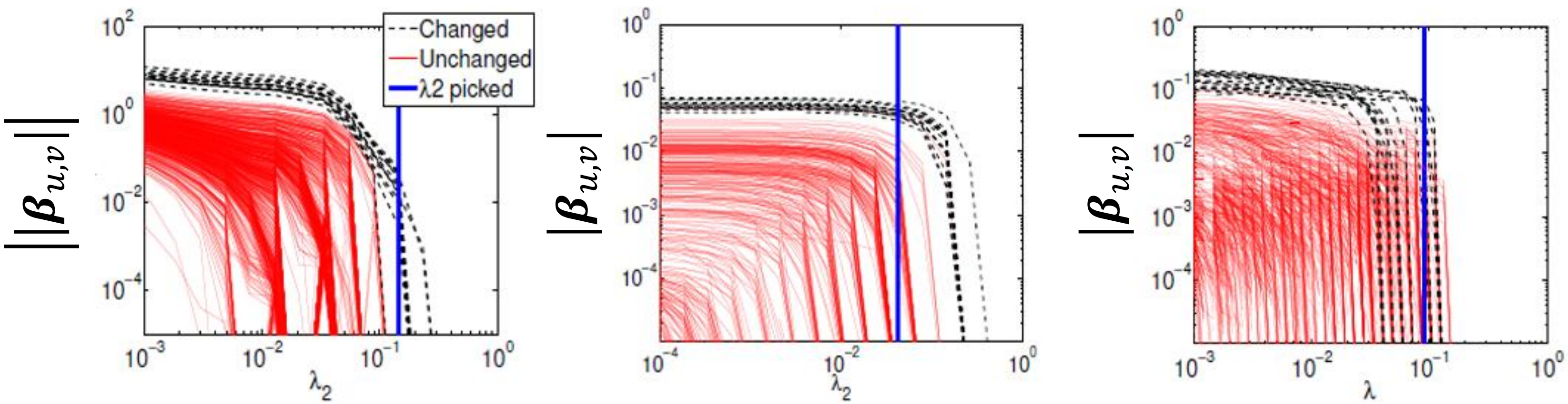


- The Illustration of methods on **model assumption** and **estimation procedure**

Chapter 3,
Structural Change Detection

1. Background
2. Existing Methods
3. Proposed Methods
4. Experiments

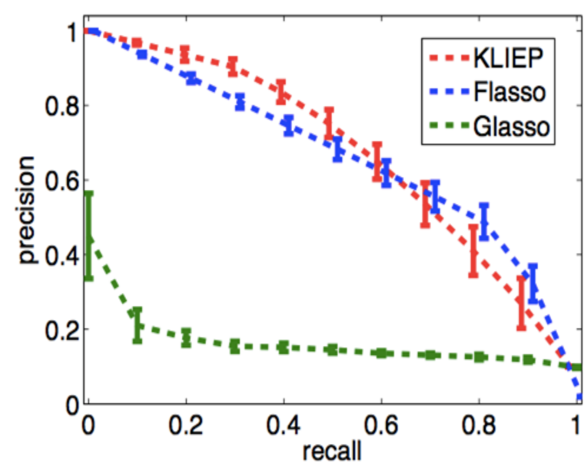
Gaussian Distribution $(n = 100, d = 40)$



(a) KLIEP, $n = 100$ (b) Flasso, $n = 100$ (c) Glasso, $n = 100$

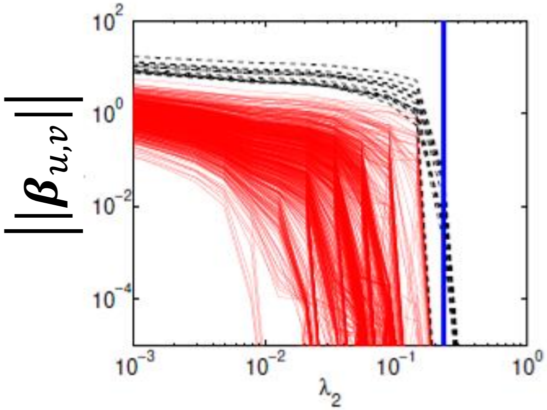
Regularization path

- Start from 40 dimensional GMN with random correlations.
- Randomly drops 15 edges.
- Precision and Recall curves are averaged over 20 runs.

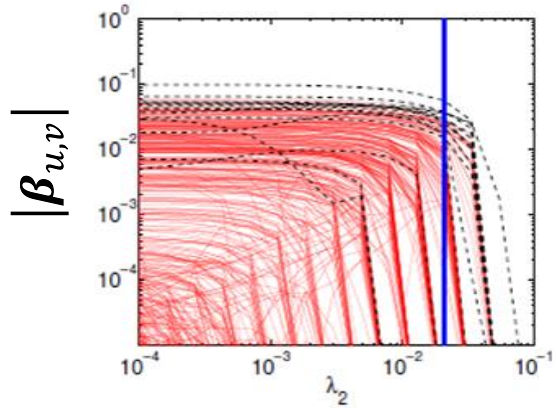


P-R curve

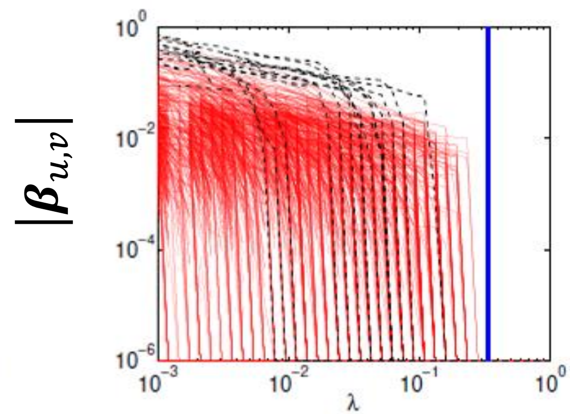
Gaussian Distribution $(n = 50, d = 40)$



(d) KLIEP, $n = 50$



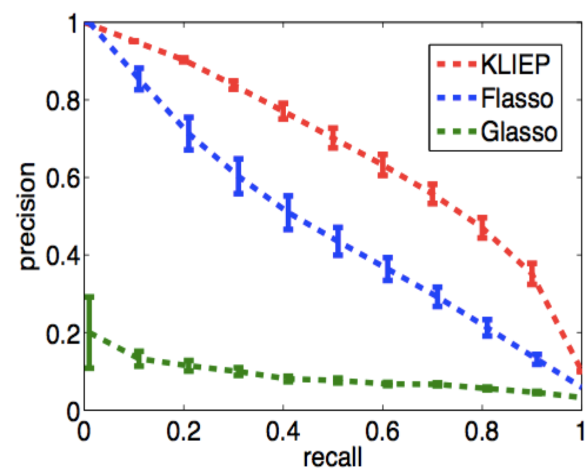
(e) Flasso, $n = 50$



(f) Glasso, $n = 50$

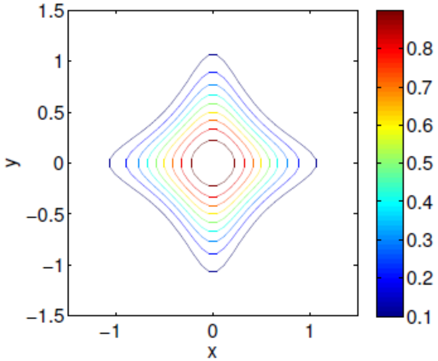
Regularization path

- Start from 40 dimensional GMN with random correlations.
- Randomly drops 15 edges.
- Precision and Recall curves are averaged over 20 runs.

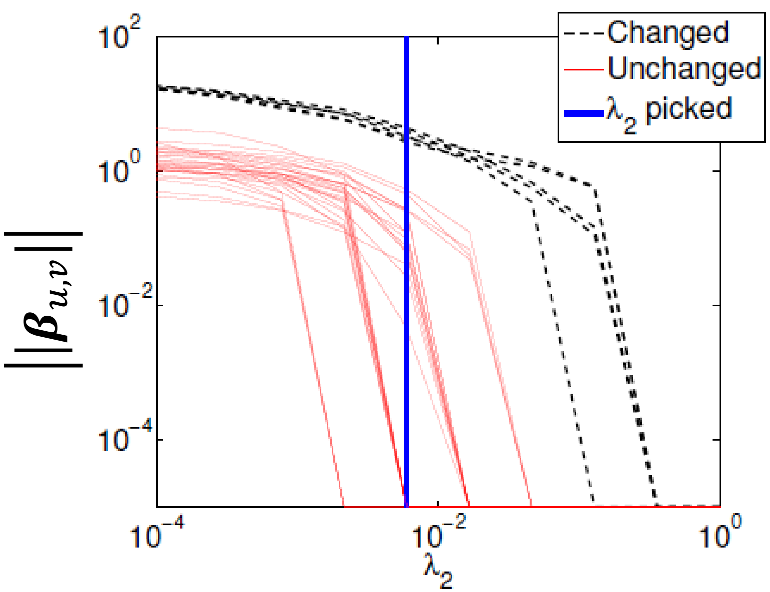


P-R curve

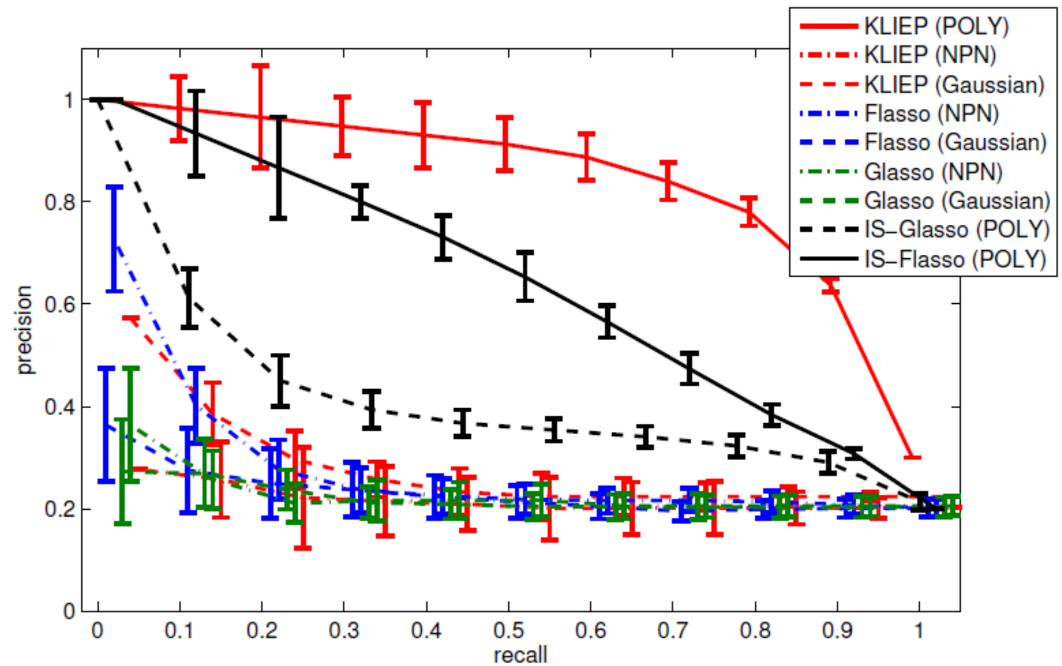
Diamond Distribution $(n = 5000, d = 9)$



$$p(\boldsymbol{x}) \propto \exp\left(-\sum_i 2\boldsymbol{x}_i^2 - \sum_{i,j\in E} 20\boldsymbol{x}_i^2\boldsymbol{x}_j^2\right)$$



Regularization path



P-R curve

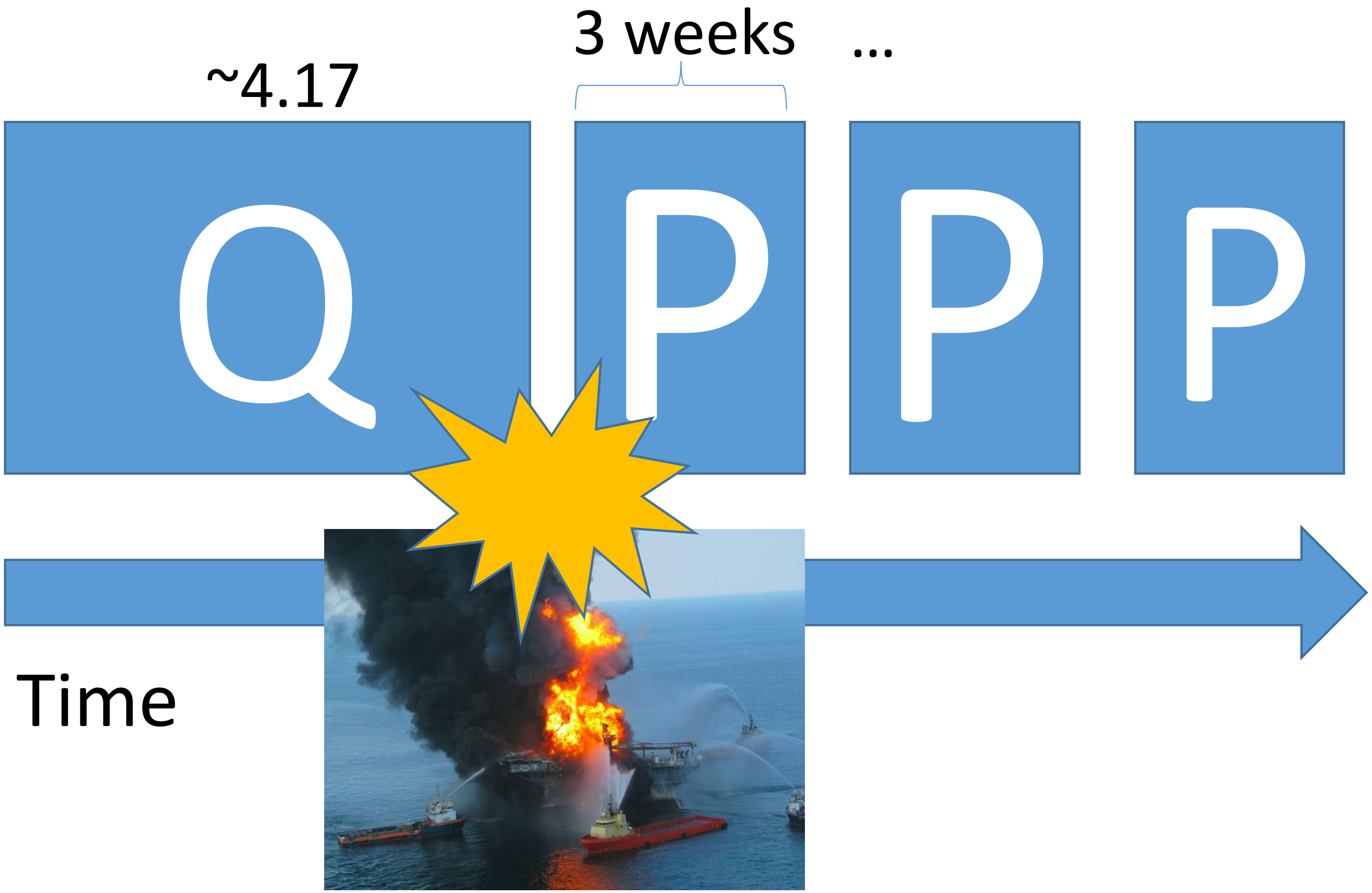
- **The proposed method has the leading performance.**

Twitter Dataset

source: Wikipedia

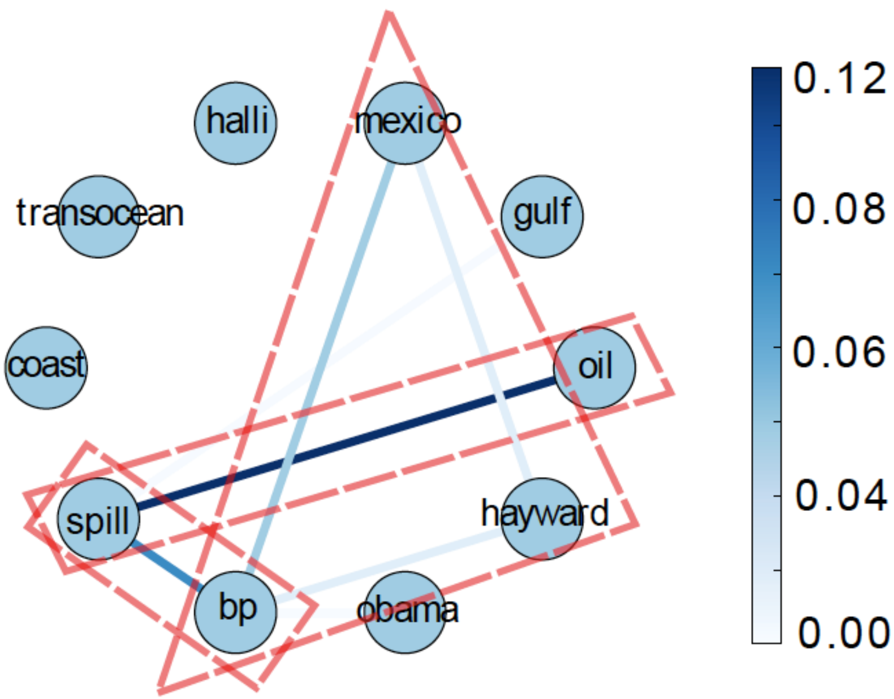
- We choose the Deepwater Horizon oil spill as the target event.
  - Samples are the frequencies of 10 related keywords over time.
  - Detecting the change of **co-occurrences on keywords** before and after a certain event.

Twitter Dataset (pp. 68-72)

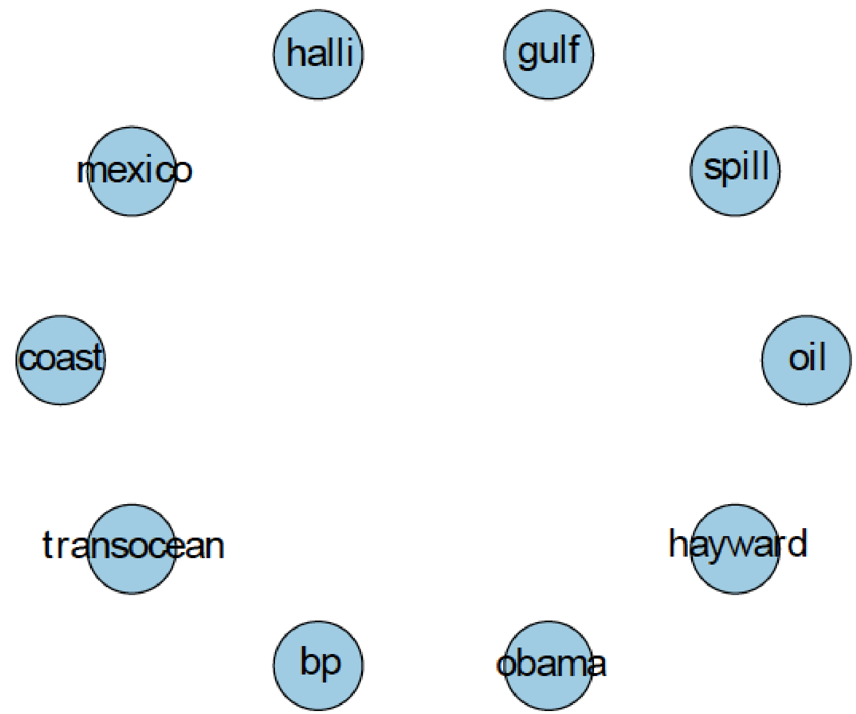


The Change of Correlation

From 7.26-9.14



KLIEP



Flasso

Small # of samples, severe over-fitting of Flasso!

Conclusion

- We proposed a **direct** structural change detection method for Markov Networks.

- The proposed method can handle **non-Gaussian data**, and can be efficiently solved using **primal** or **dual** objective.

- The proposed method showed the **leading** performance on both artificial and real-world datasets