MLSS, Beijing Jun. 16, 2014

Unsupervised Change Detection

Masashi Sugiyama

Tokyo Institute of Technology, Japan

sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi/

Abstract

We consider the problem of detecting change in two sets of samples, and explore two approaches: distributional and structural change detection.

Distributional change detection is aimed at estimating a divergence between the probability densities behind the two sets of samples. We first explain that the two-step approach of first estimating the probability densities and then computing the divergence from the estimated densities results in systematic under-estimation of the divergence. Then we introduce methods to direct estimate the ratio of densities and the difference of densities, which are shown to be more reliable than the density estimation approach.

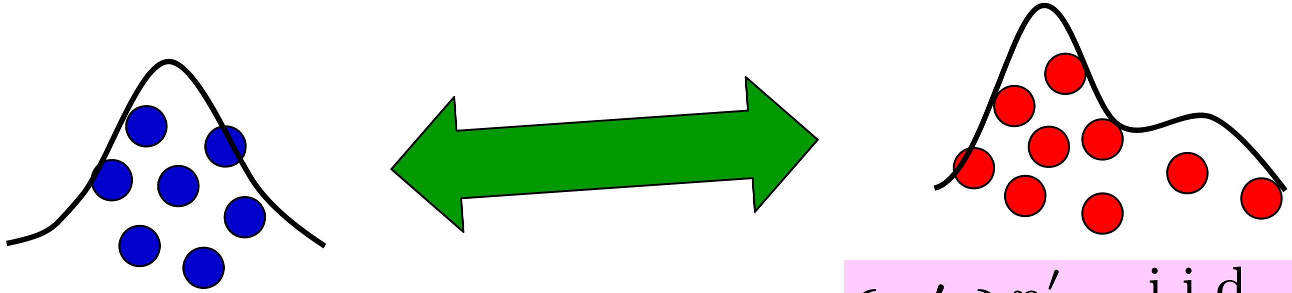
Abstract (cont.)

Structural change detection tries to identify change in element-wise dependency structure in multi-dimensional samples. We first consider the Gaussian sparse covariance selection setup and introduce approaches based on LASSO and fused-LASSO. Then we extend our discussion to non-Gaussian Markov networks, which generally suffer computational intractability of the normalization term, and introduce the importance sampling technique and the score matching method. Finally, we cover a method to directly compare two Markov networks for change detection.

No solid background on change detection is necessary, but basic knowledge of elementary statistics, linear algebra, and optimization is assumed.

Change Detection

- Goal: Given two sets of samples, we want to compare the probability distributions behind



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

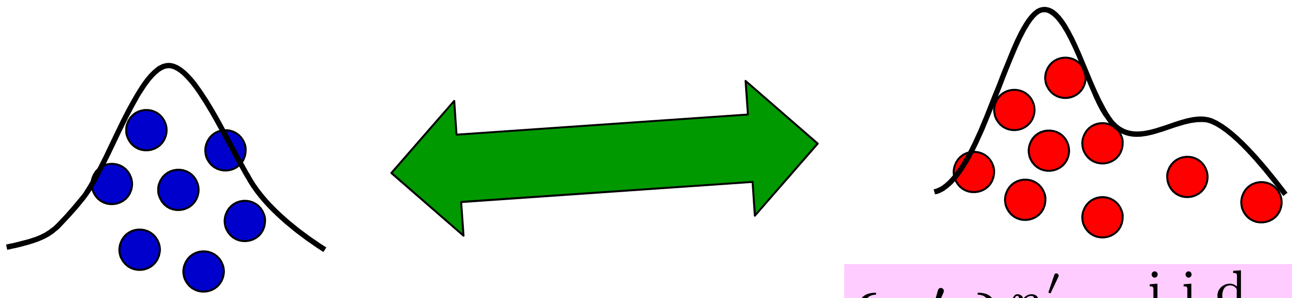
- Two approaches:
  - Distributional change detection: Flexible and robust
  - Structural change detection: Interpretable

Contents

1. Distributional change detection
   - A) Problem setup and motivating examples
   - B) Distances
   - C) Distance approximation
2. Structural change detection

Distributional Change Detection

- Goal: Detect change in probability distributions behind two sets of samples through divergence



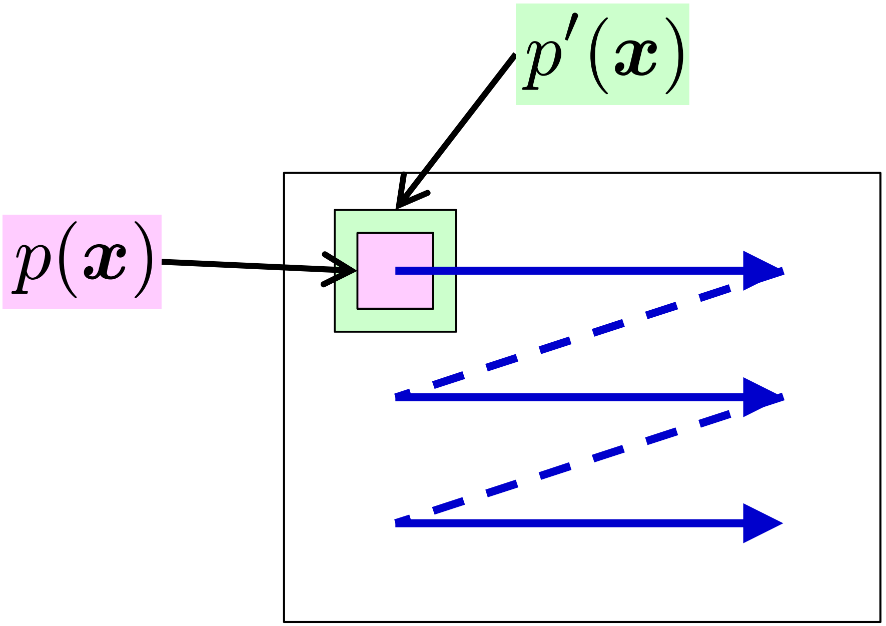
$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

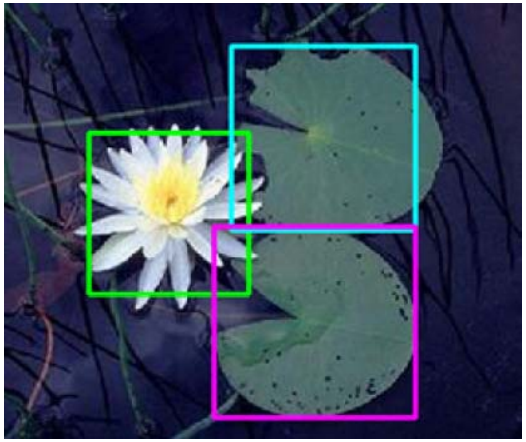
$$\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$$

$$\mathrm{Divergence}(p, p') < \varepsilon \ ?$$

Motivating Example 1

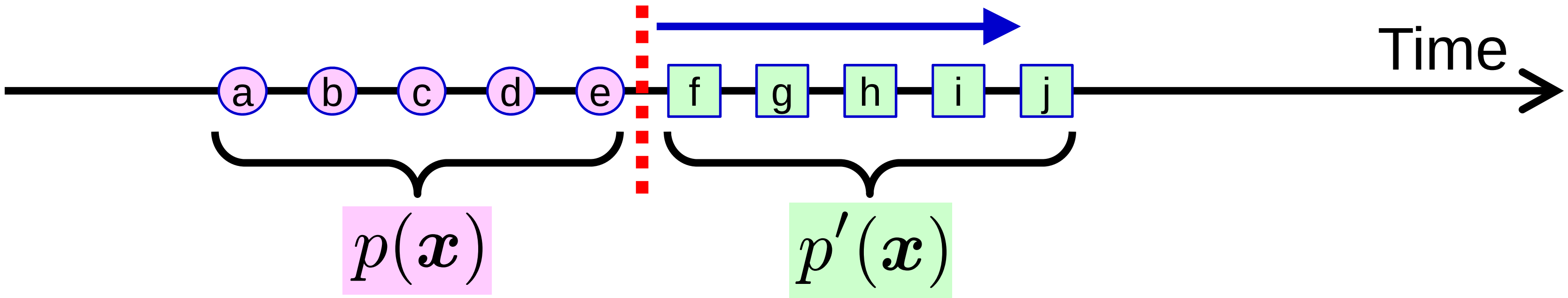
- Region-of-interest detection in images:
  - $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are significantly different when a visually salient object is included inside.

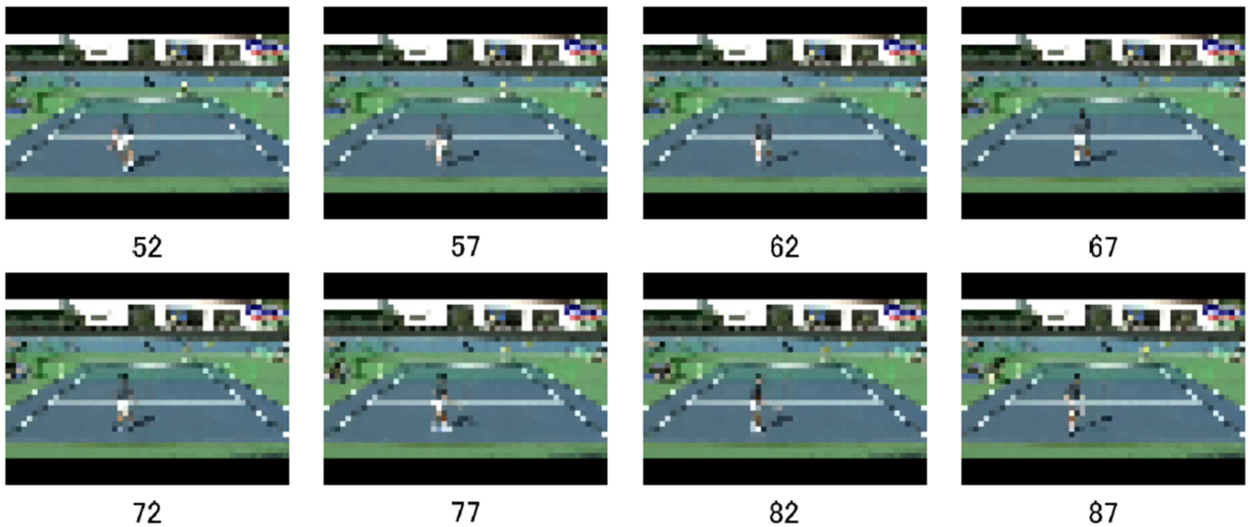




Motivating Example 2

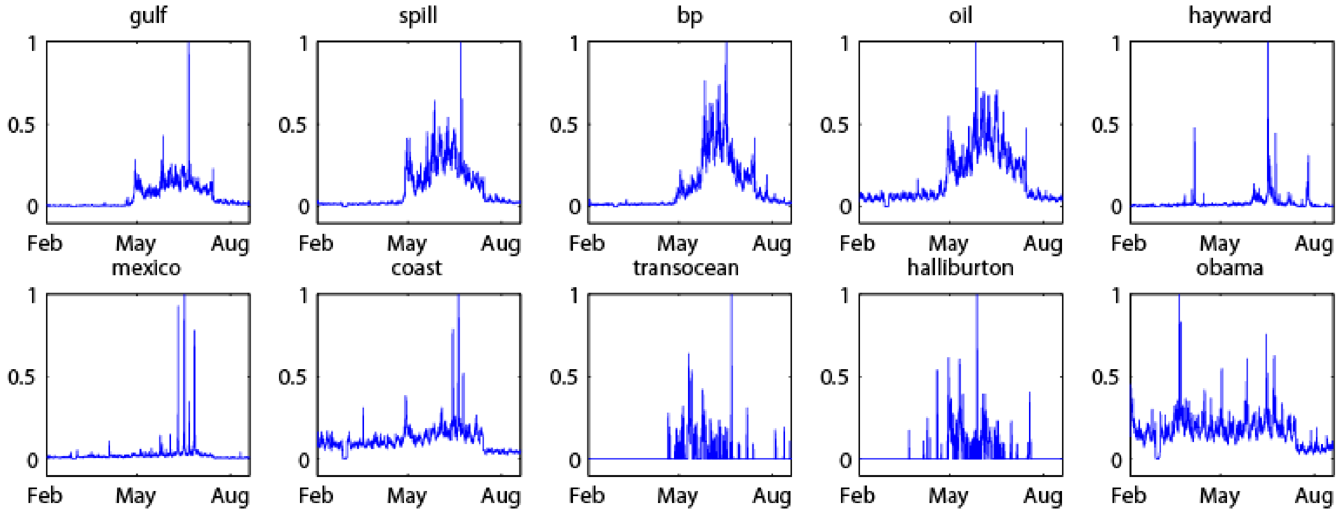
- Event detection in movies:
  - $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are significantly different when an irregular event occurs.

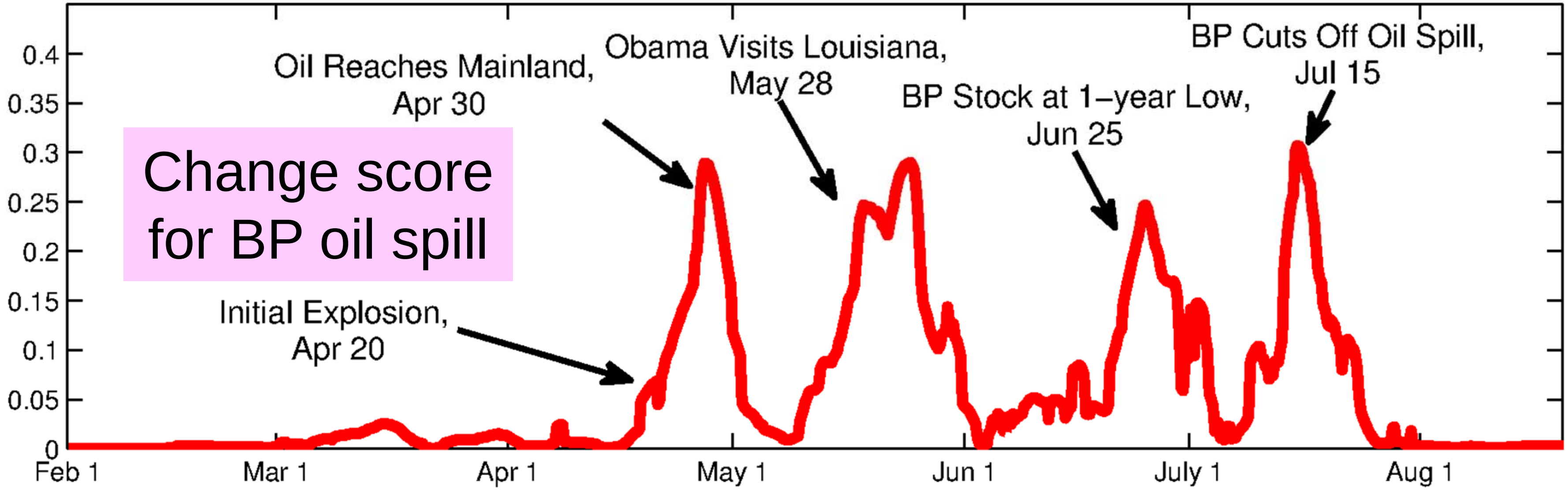




Motivating Example 3

- Event detection from Twitter:



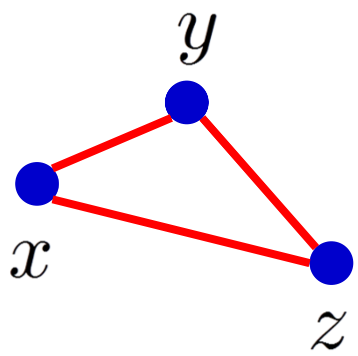


Contents

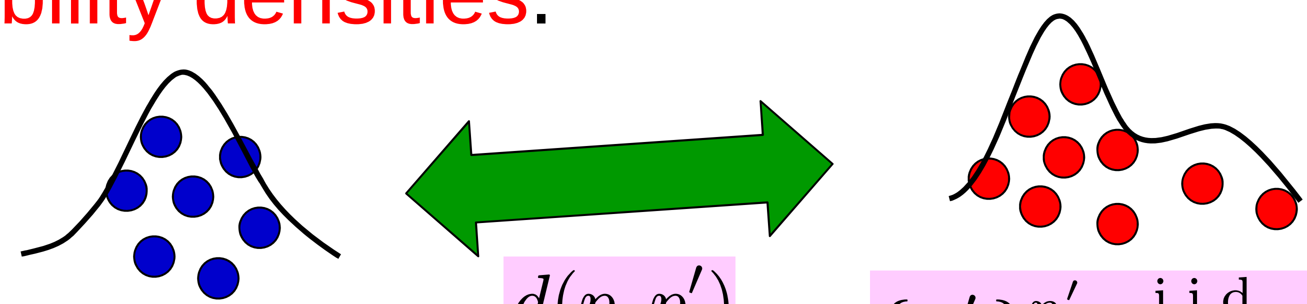
Distances and Divergences

Distance:

- Non-negativity: $\forall x, y, \quad d(x,y) \geq 0$
- Non-degeneracy: $d(x,y) = 0 \iff x = y$
- Symmetry: $\forall x, y, \quad d(x,y) = d(y,x)$
- Triangularity: $\forall x, y, z \quad d(x,z) \leq d(x,y) + d(y,z)$



- A divergence is a pseudo-distance.
- We consider distances/divergences between probability densities.



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

$d(p, p')$

$\{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$

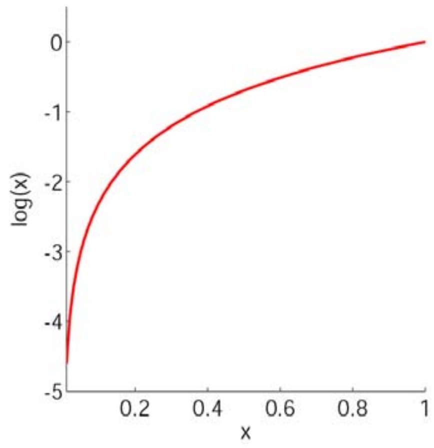
Contents

Kullback-Leibler Divergence

Kullback & Leibler (1951)

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

- ☺ Compatible with maximum likelihood.
- ☺ Invariant under input transformation. (Jacobians cancel in the density ratio) $\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$
- ☹ Doesn’t satisfy symmetry and triangularity.
- ☹ Sensitive to outliers (due to log and ratio).



f-Divergences

Ali & Silvey (1966), Csiszár (1967)

$$F(p\|p') = \int p'(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$$

f : Convex function such that $f(1) = 0$

- $f(t) = t \log t$ yields the KL-divergence:

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

- To avoid the log function, let us use

$$f(t) = (t-1)^2$$

Pearson (PE) Divergence

Pearson (1900)

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

- ☺ Compatible with least-squares.
- ☺ Invariant under input transformation.
- ☹ Doesn't satisfy symmetry and triangularity.
- ☹ Sensitive to outliers (no log, but still ratio).

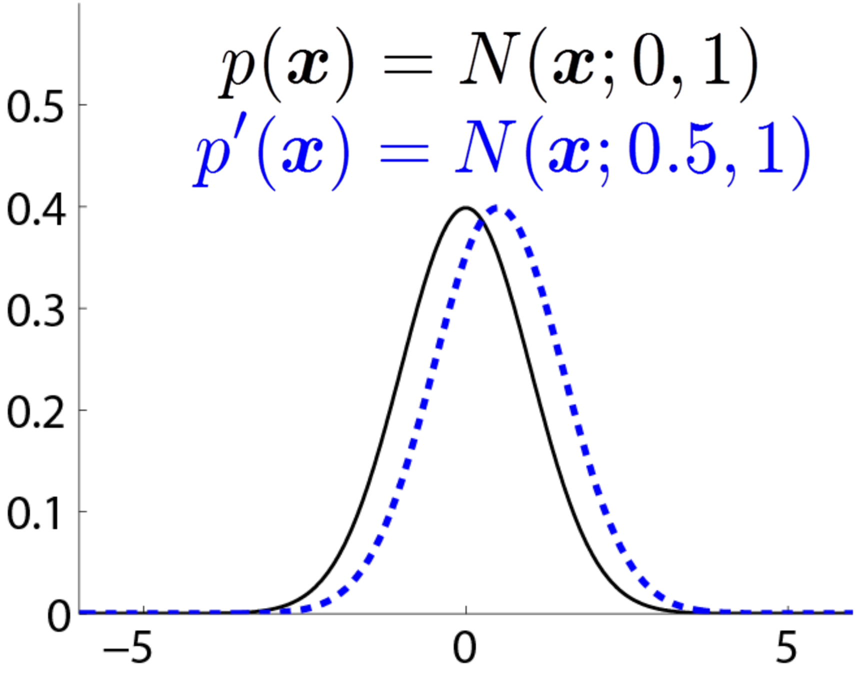
$$\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$$

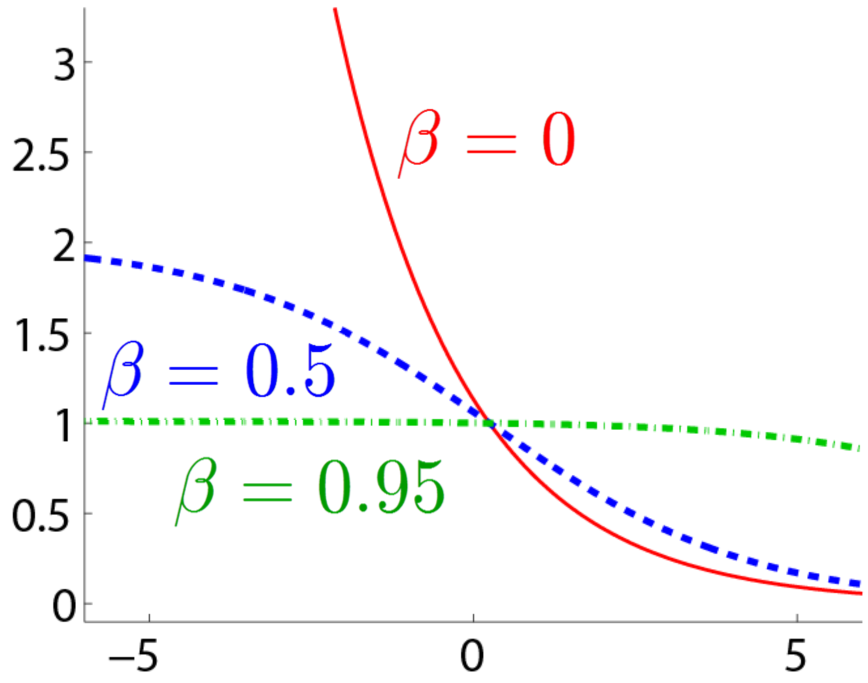
Relative Density Ratio

Yamada *et al.* (NIPS2011, NeCo2013)

- Density ratio $\dfrac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$ can diverge to infinity.
- Relative density ratio is always bounded:

$$\frac{p(\boldsymbol{x})}{\beta p(\boldsymbol{x}) + (1-\beta)p'(\boldsymbol{x})} < \frac{1}{\beta} \qquad 0 \le \beta < 1$$





Relative Pearson (rPE) Divergence

$$\mathrm{rPE}(p\|p') = \mathrm{PE}(p\|p_\beta) = \int p_\beta(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{p_\beta(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$$

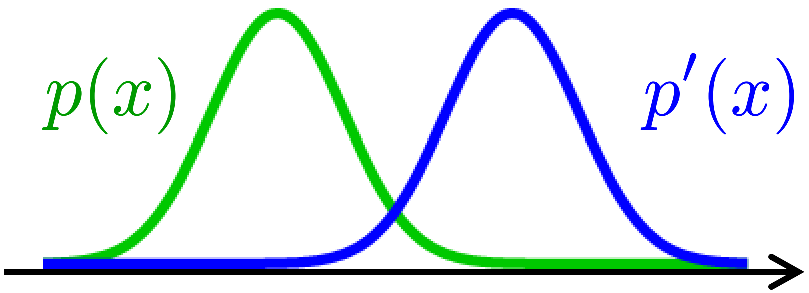
$0 \leq \beta < 1$ $p_\beta(\boldsymbol{x}) = \beta p(\boldsymbol{x}) + (1-\beta)p'(\boldsymbol{x})$

- ☺ Compatible with least-squares.
- ☺ Invariant under input transformation.
- ☺ Robust against outliers.
- ☹ Doesn't satisfy symmetry and triangularity.
- ☹ Not clear how to choose β.

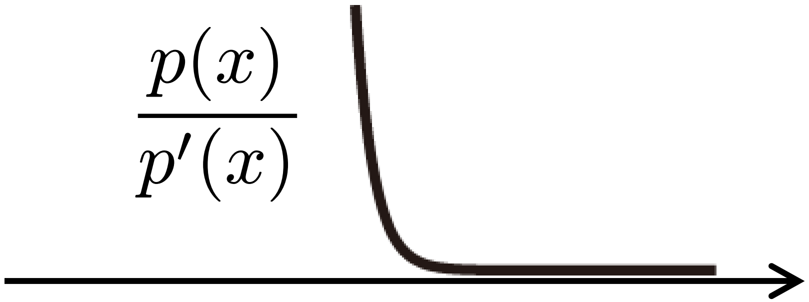
Contents

1. Distributional change detection
   - A) Problem setup and motivating examples
   - B) Distances
     - I. Density-ratio divergences
     - II. Density-difference distances
   - C) Distance approximation
2. Structural change detection

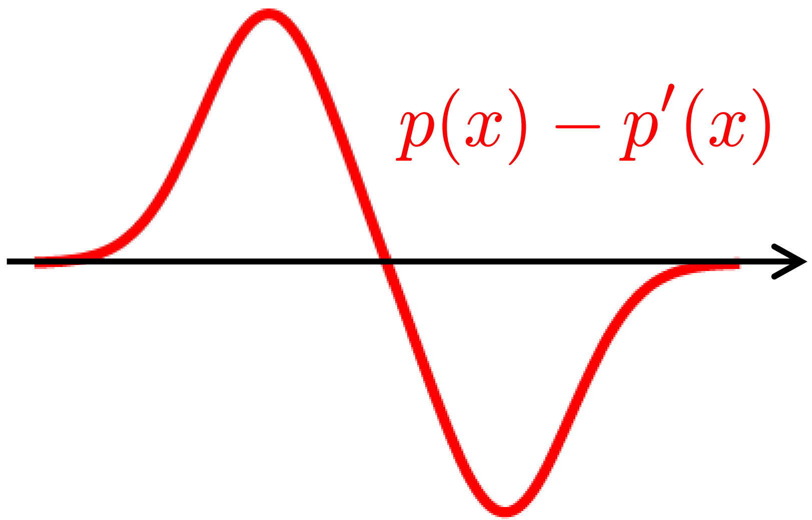
Density Difference



- Density ratio based distance:
  - Is the ratio 1?



- Density difference based distance:
  - Is the difference 0?



L^t-Distance

$$L^t(p,p') = \int \left| p(\boldsymbol{x}) - p'(\boldsymbol{x}) \right|^t \mathrm{d}\boldsymbol{x} \qquad t \geq 0$$

- ☺ Proper distance.
- ☺ Robust against outliers (no ratio).

■ When $t = 2$:

- ☺ Compatible with least-squares.
- ☹ Not invariant under input transformation.

■ When $t = 1$:

- ☺ Invariant under input transformation (because f-div).

$$f(t) = |t-1|$$

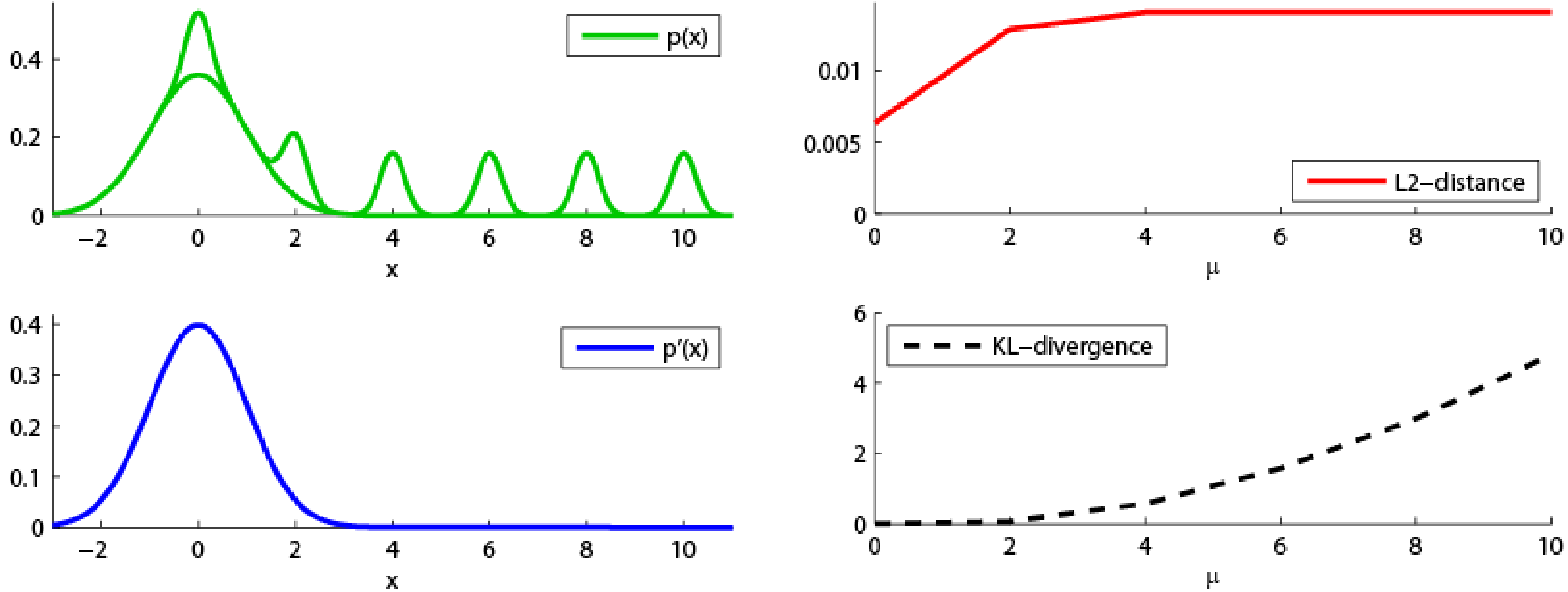
$$L^1(p,p') = \int p'(\boldsymbol{x}) \left| \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} - 1 \right| \mathrm{d}\boldsymbol{x}$$

KL vs. L^2 for Outliers

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

$$L^2(p, p') = \int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}$$

$$p(x) = 0.9p'(x) + 0.1q(x - \mu)$$



- L^2-distance is bounded.
- KL-divergence is unbounded.

Contents

1. Distributional change detection
   A) Problem setup and motivating examples
   B) Distances
   C) Distance approximation
      I. Density-ratio divergences
      II. Density-difference distances
2. Structural change detection

Distance Approximation via Density Estimation

1. Estimate densities $p(\boldsymbol{x}), p'(\boldsymbol{x})$ from samples:

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

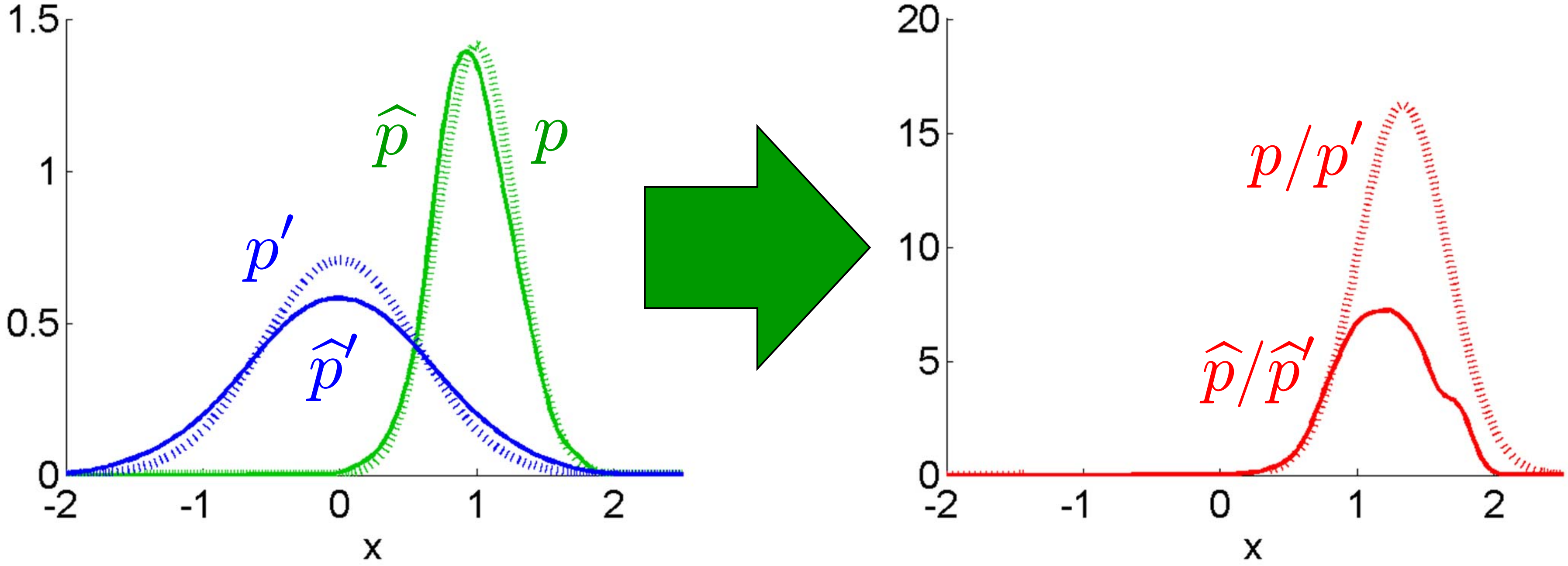
   - Maximum likelihood estimation
   - Bayes estimation
   - Kernel density estimation
   - Nearest-neighbor density estimation.

2. Plug-in the estimated densities $\widehat{p}(\boldsymbol{x}), \widehat{p'}(\boldsymbol{x})$:

$$\widehat{\mathrm{KL}}(p\|p') = \int \widehat{p}(\boldsymbol{x}) \log \frac{\widehat{p}(\boldsymbol{x})}{\widehat{p'}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \qquad \widehat{L}^2(p, p') = \int \left(\widehat{p}(\boldsymbol{x}) - \widehat{p'}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}$$

Drawback of Plug-In Density Estimation Approach

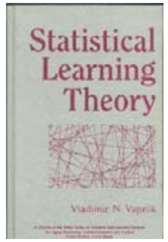
- Densities are estimated without regard to taking their ratio later.
- Division by $\widehat{p}'$ magnifies estimation error in $\widehat{p}$.



Guiding Principle

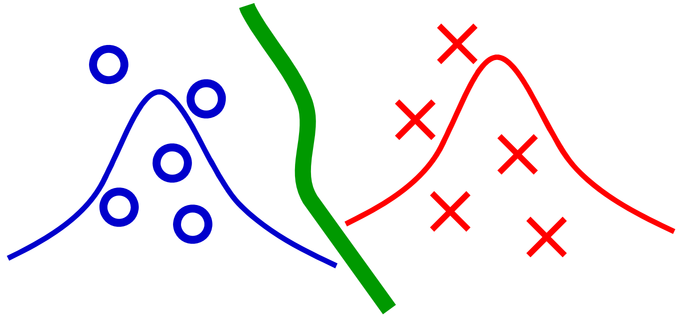
- Vapnik's principle: Vapnik (Wiley 1998)

When solving a problem of interest, one should not solve a more general problem as an intermediate step



  - Support vector machine avoids general density estimation and directly learns the boundary.

    Cortes & Vapnik (MLJ1995)



- Let's avoid separately estimating $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$, and directly compare the densities!

Vapnik's Principle in Distance Approximation

$$\mathrm{KL}(p\|p') = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \qquad L^2(p,p') = \int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}$$

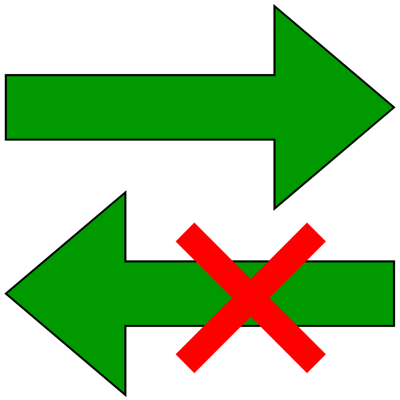
- Directly estimate the density ratio / difference:

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \qquad f(\boldsymbol{x}) = p(\boldsymbol{x}) - p'(\boldsymbol{x})$$

without estimating each density $p(\boldsymbol{x}), p'(\boldsymbol{x})$.

Knowing densities

$p(\boldsymbol{x}), p'(\boldsymbol{x})$



Knowing density ratio / difference

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \qquad f(\boldsymbol{x}) = p(\boldsymbol{x}) - p'(\boldsymbol{x})$$

Contents

1. Distributional change detection
   - A) Problem setup and motivating examples
   - B) Distances
   - C) Distance approximation
     - I. Density-ratio divergences
     - II. Density-difference distances
2. Structural change detection

KL-Divergence Approximation

Nguyen *et al.* (NIPS2007, IEEE-IT2010)
Sugiyama *et al.* (NIPS2007, AISM2008)

$$\mathrm{KL}(p\|p') = \int p(\boldsymbol{x}) \log r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \qquad r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$$

- Directly approximate the density ratio with log-loss:

$$\widehat{r} = \underset{\widetilde{r}}{\mathrm{argmin}}\ \mathrm{KL}(p\|\widetilde{r}\cdot p')$$

$$\text{subject to} \quad \int \widetilde{r}(\boldsymbol{x}) p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \quad \text{and} \quad \widetilde{r} \geq 0$$

$$\Rightarrow \quad \mathrm{KL}(p\|p') \approx \int p(\boldsymbol{x}) \log \widehat{r}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

- Expectation is approximated by empirical average.

Solution for Linear Model

- Linear-in-parameter model:

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{b} \alpha_j \phi_j(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top$:Fixed basis functions

$\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_b(\boldsymbol{x}))^\top$:Parameters

- Empirical optimization problem:

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\mathrm{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log r_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)$$

$$\text{subject to } \frac{1}{n'} \sum_{i'=1}^{n'} r_{\boldsymbol{\alpha}}(\boldsymbol{x}'_{i'}) = 1 \text{ and } \boldsymbol{\alpha} \geq \mathbf{0}$$

  - The solution tends to be sparse due to $\boldsymbol{\alpha} \geq \mathbf{0}$.

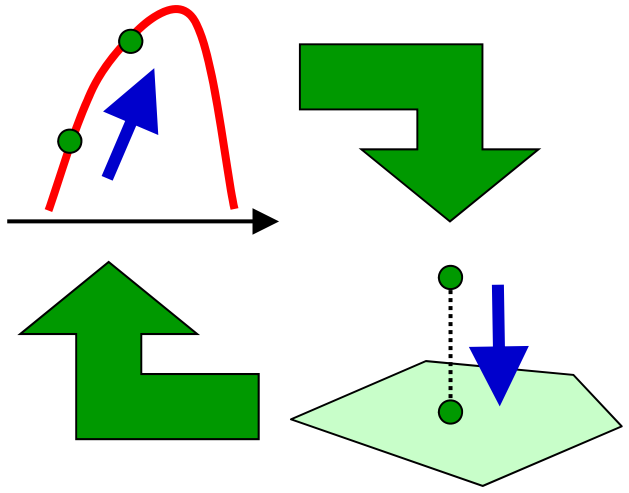
Solution for Linear Model

$$\widehat{\boldsymbol{\alpha}} = \mathop{\mathrm{argmax}}_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^{n} \log \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}_i)$$

$$\text{subject to } \frac{1}{n'}\sum_{i'=1}^{n'} \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}'_{i'}) \text{ and } \boldsymbol{\alpha} \geq \mathbf{0}$$

- Thanks to convexity, global optimal solution can be obtained by simple gradient-projection.
- Resulting KL-divergence approximator:

$$\mathrm{KL}(p\|p') \approx \frac{1}{n}\sum_{i=1}^{n} \log \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(\boldsymbol{x}_i)$$



Other Models

- **Kernel model:**

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

  - ☺ Nonparametric

- **Log-linear model:**

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \exp\left(\sum_{j=1}^{b} \alpha_j \phi_j(\boldsymbol{x})\right)$$

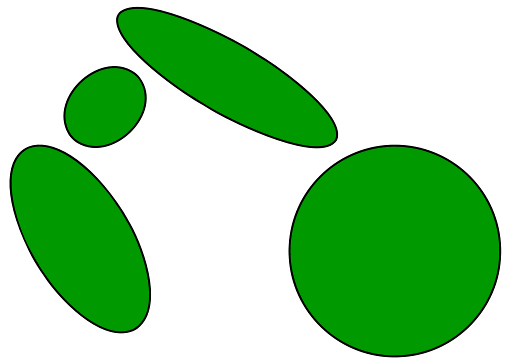
  - ☺ Always positive
  - ☺ Compatible with Markov networks

- **Gaussian mixture model:**

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{b} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

  - ☺ More flexible
  - ☹ Non-convex optimization

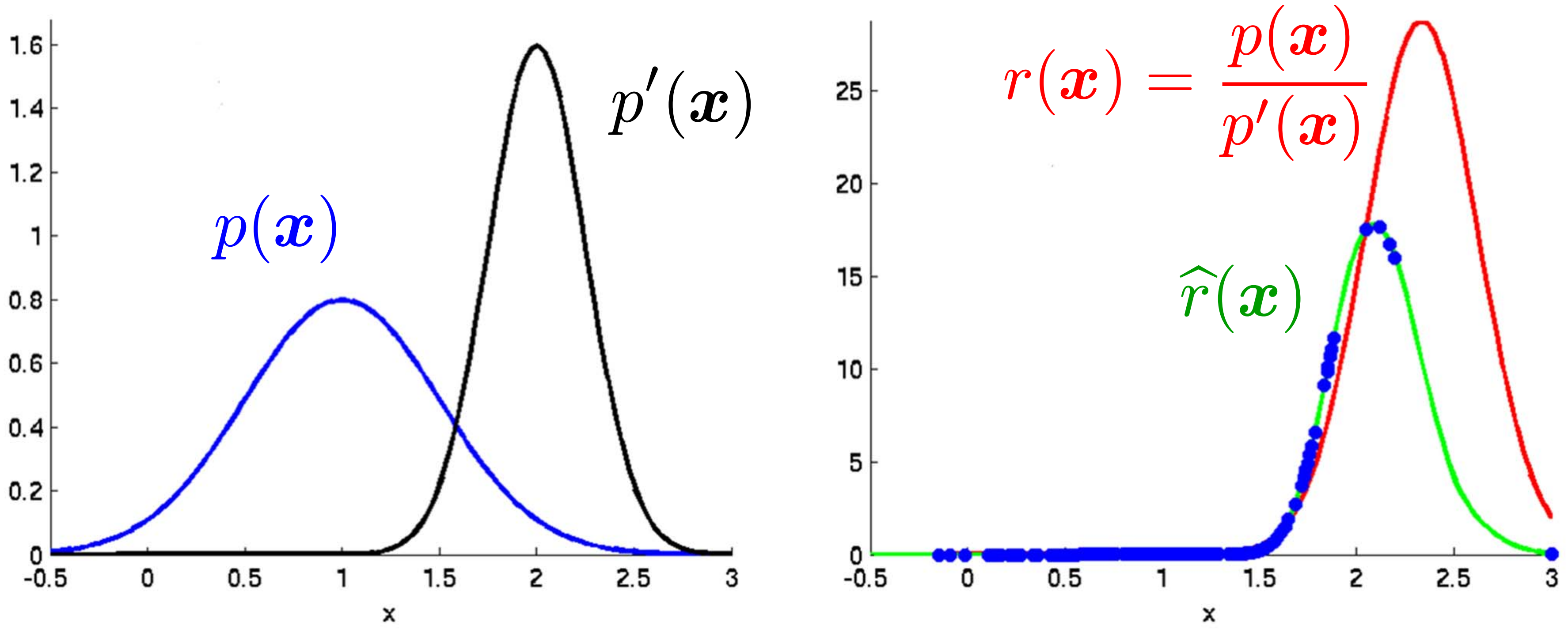
- **Probabilistic PCA mixture:**
  - ☺ Local dimension reduction
  - ☹ Non-convex optimization



Numerical Example

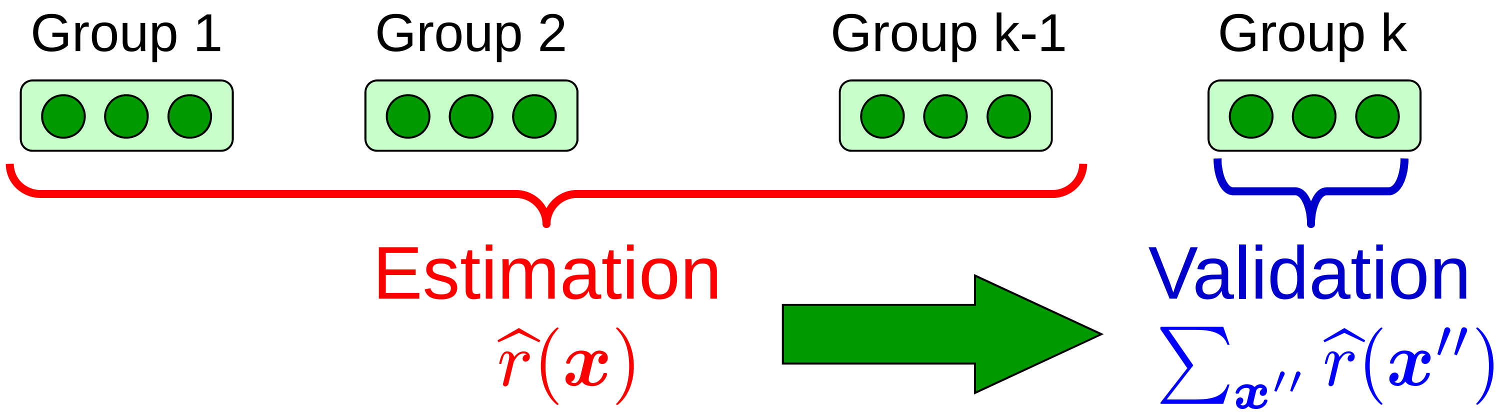
- Gaussian kernel model:

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$$



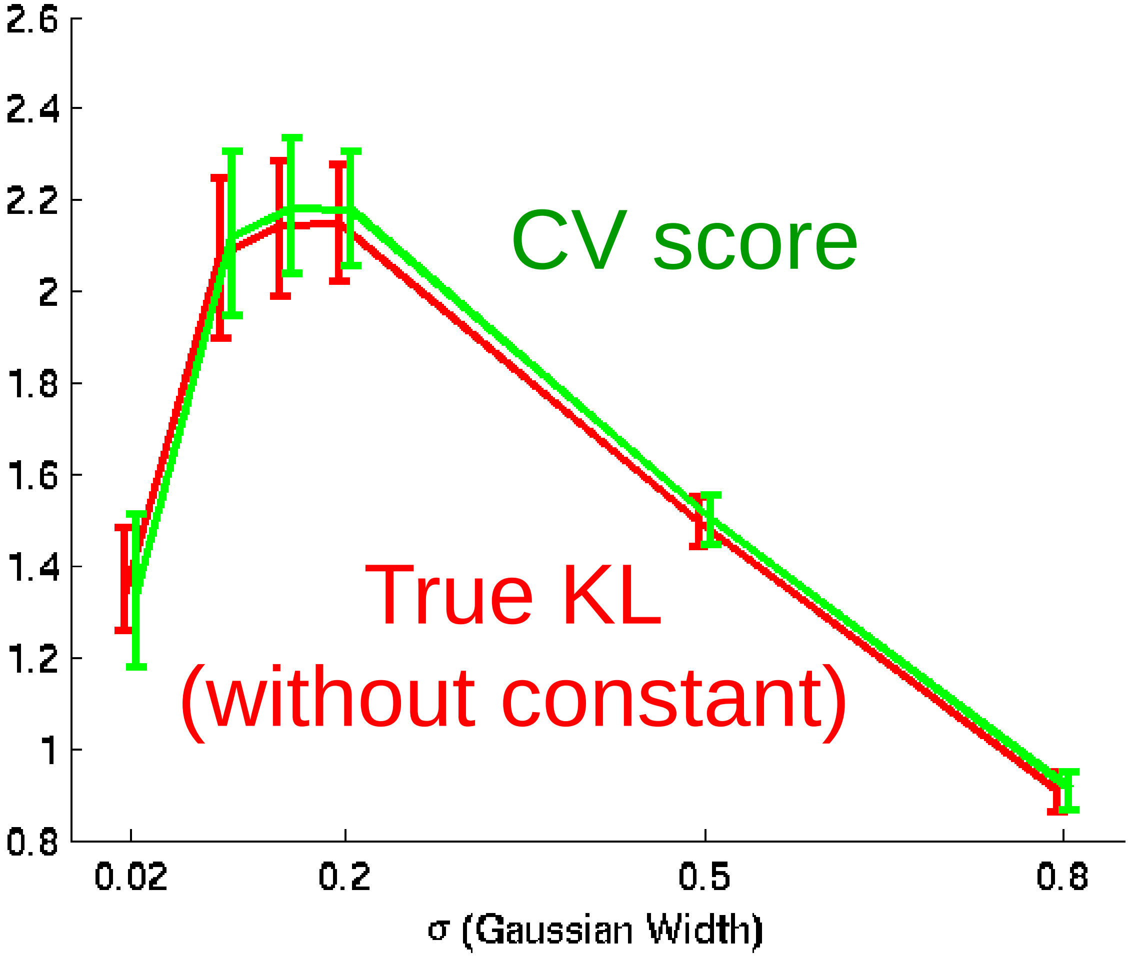
Model Selection

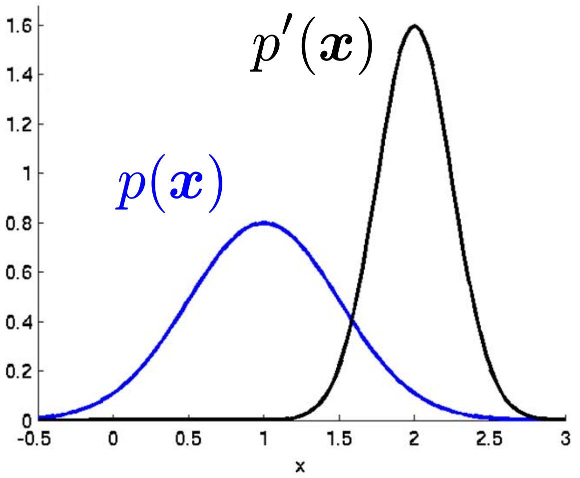
- Choice of the Gaussian bandwidth affects the performance.
- Cross-validation (CV):
  - Split $\{\boldsymbol{x}_i\}_{i=1}^{n}$ into estimation and validation subsets.

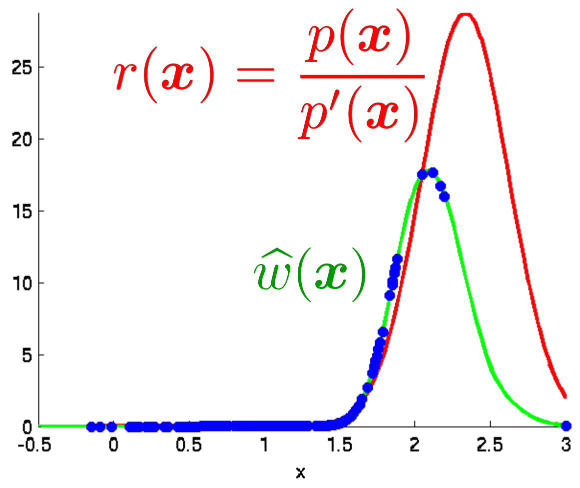


  - Repeat this estimation-validation process for all combinations
- CV gives an almost unbiased estimator of KL.

Numerical Example



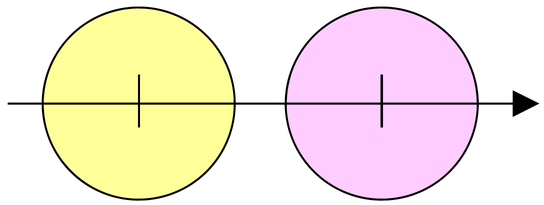




- Model selection by CV works.

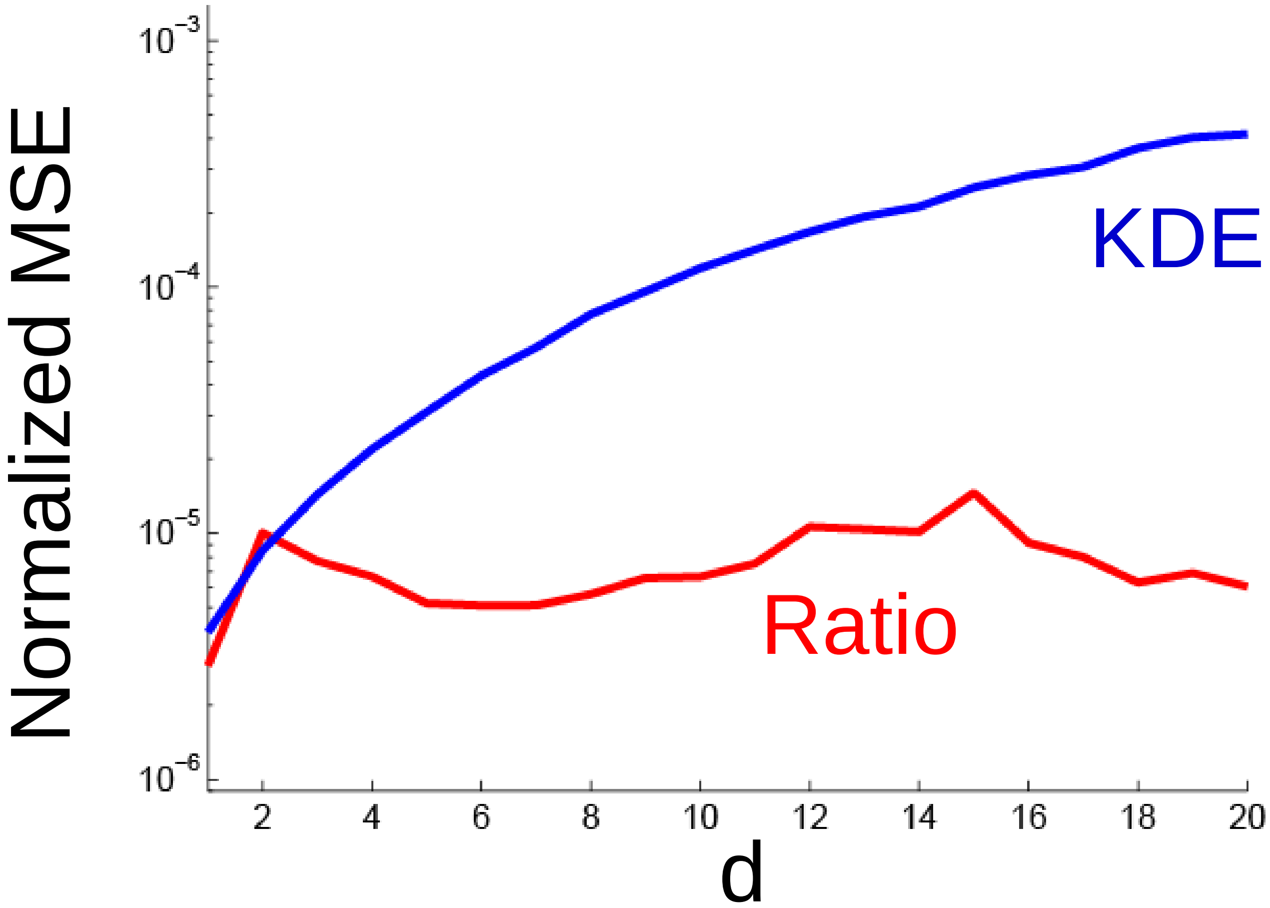
Comparison with KDE

- d-dimensional Gaussians with covariance identity and
  - Denominator: mean (0,0,0,…,0)
  - Numerator: mean (1,0,0,…,0)



- Kernel density estimation (KDE):
  - Estimate two densities separately and take ratio.
  - Gaussian widths are chosen by CV.
- Ratio:
  - Estimate the density ratio directly.
  - Gaussian width is chosen by CV.

Accuracy as a Function of Input Dimensionality



- Density ratio approach works better.

PE-Divergence Approximation

Kanamori *et al.* (NIPS2008, JMLR2009)

$$\mathrm{PE}(p\|p') = \int p'(\boldsymbol{x})\Big(r(\boldsymbol{x}) - 1\Big)^2 \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x})r(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - 1$$

- Directly approximate the density ratio by least-squares:

$$\widehat{r} = \underset{\widetilde{r}}{\mathrm{argmin}} \int p'(\boldsymbol{x})\Big(\widetilde{r}(\boldsymbol{x}) - r(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}$$

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$$

$$= \underset{\widetilde{r}}{\mathrm{argmin}} \int p'(\boldsymbol{x})\Big(\widetilde{r}(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x} - 2\int p(\boldsymbol{x})r(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

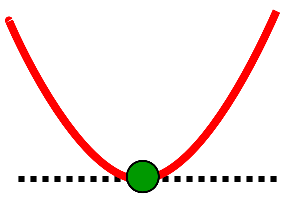
$$\Longrightarrow \quad \mathrm{PE}(p\|p') \approx \int p(\boldsymbol{x})\widehat{r}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - 1$$

- Expectation is approximated by empirical average.

PE-Divergence Approximation for Linear Model

$$\widehat{\boldsymbol{\alpha}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}} \left[\frac{1}{n'} \sum_{i'=1}^{n'} r_{\boldsymbol{\alpha}}(\boldsymbol{x}'_{i'})^2 - \frac{2}{n} \sum_{i=1}^{n} r_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) \right]$$

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$



- Solution is given analytically:

$$\widehat{\boldsymbol{G}} = \frac{1}{n'} \sum_{i'=1}^{n'} \boldsymbol{\phi}(\boldsymbol{x}'_{i'}) \boldsymbol{\phi}(\boldsymbol{x}'_{i'})^\top$$

$$\widehat{\boldsymbol{h}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)$$

$$\widehat{\boldsymbol{\alpha}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}} \left[\boldsymbol{\alpha}^\top \widehat{\boldsymbol{G}} \boldsymbol{\alpha} - 2 \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right]$$

$$= (\widehat{\boldsymbol{G}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$$

- Resulting PE-divergence approximator:

$$\mathrm{PE}(p \| p') \approx \frac{1}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - 1 = \widehat{\boldsymbol{h}}^\top (\widehat{\boldsymbol{G}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}} - 1$$

MATLAB Implementation for Gauss Kernel Model

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$$

$$\mathrm{PE}(p\|p') \approx \widehat{\boldsymbol{h}}^\top (\widehat{\boldsymbol{G}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}} - 1$$

$$\widehat{h}_j = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$$

$$\widehat{G}_{j,j'} = \frac{1}{n'}\sum_{i'=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x}'_{i'}-\boldsymbol{x}_j\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{x}'_{i'}-\boldsymbol{x}_{j'}\|^2}{2\sigma^2}\right)$$

```
n=1000; x=randn(n,1); y=randn(n,1)+1/2;
x2=x.^2; xx=repmat(x2,1,n)+repmat(x2',n,1)-2*x*x'; s=exp(-xx);
y2=y.^2; yx=repmat(y2,1,n)+repmat(x2',n,1)-2*y*x'; t=exp(-yx);
PE=mean(s*((t'*t/n+eye(n))¥(mean(s,2))))-1;
```

- Relative density ratio can also be estimated in the almost same way.

f-Divergences and Duality

$$F(p\|p') = \int p'(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$$

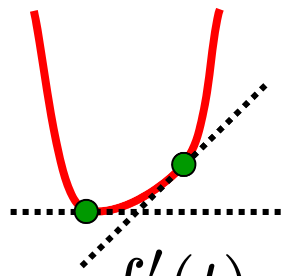
f : Convex function such that $f(1) = 0$

- Fenchel transform (convex conjugate):

$$f^*(r) = -\inf_t [f(t) - rt]$$

- Conjugate of conjugate:

$$f(t) = -\inf_r [f^*(r) - rt]$$



- KL-divergence: $f(t) = t \log t$ $\quad f^*(r) = \exp(r-1)$
- PE-divergence: $f(t) = (t-1)^2$ $\quad f^*(r) = r^2/2 + r$

Lower Bound of f-Divergences

Nguyen *et al.* (NIPS2007, IEEE-IT2010)

$$F(p\|p') = \int p'(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$$

$$f(t) = -\inf_{r}[f^*(r) - rt]$$

- Lower bound of f-divergences:

$$F(p\|p') = -\inf_{r}\left[\int p'(\boldsymbol{x}) f^*(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})r(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right]$$

- Sample approximation gives

$$\widehat{F}(p\|p') = -\min_{\boldsymbol{\alpha}}\left[\frac{1}{n'}\sum_{i'=1}^{n'} f^*(r_{\boldsymbol{\alpha}}(\boldsymbol{x}'_{i'})) - \frac{1}{n}\sum_{i=1}^{n} r_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)\right]$$

Contents

L^2-Distance Approximation

Kim & Scott (IEEE-TPAMI2010)
Sugiyama *et al.* (NIPS2012, NeCo2013)

$$L^2(p,p') = \int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} \qquad f(\boldsymbol{x}) = p(\boldsymbol{x}) - p'(\boldsymbol{x})$$

- Directly approximate the density difference by LS:

$$\widehat{f} = \underset{\widetilde{f}}{\mathrm{argmin}} \int \left(\widetilde{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}$$

$$= \underset{\widetilde{f}}{\mathrm{argmin}} \int \left(\widetilde{f}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} - 2\int f(\boldsymbol{x})\widetilde{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\Rightarrow \quad L^2(p,p') \approx \int \widehat{f}(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$$

- Expectation is approximated by empirical average.

Solution for Linear Model

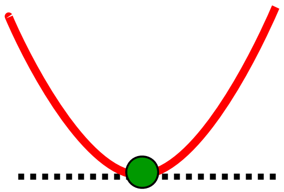
$$\widehat{\boldsymbol{\alpha}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}} \int \left(f_{\boldsymbol{\alpha}}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} + \frac{1}{n'}\sum_{i'=1}^{n'} f_{\boldsymbol{\alpha}}(\boldsymbol{x}'_{i'})^2 - \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)$$

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{b} \alpha_j \phi_j(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

- (Regularized) solution is given analytically:

$$\widehat{\boldsymbol{\alpha}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}} \left[\boldsymbol{\alpha}^\top \boldsymbol{G}\boldsymbol{\alpha} - 2\widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right]$$

$$= (\boldsymbol{G} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$$



$$\boldsymbol{G} = \int \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x} \qquad \widehat{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{i'=1}^{n'} \boldsymbol{\phi}(\boldsymbol{x}'_{i'})$$

Resulting L^2-Distance Approximator

- Two ways to approximate the L^2-distance by density-difference estimation:
  - $$L^2(p,p') = \int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} \approx \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{G}\widehat{\boldsymbol{\alpha}}$$
    $$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{G} + \lambda \boldsymbol{I})^{-1}\widehat{\boldsymbol{h}} \quad f(\boldsymbol{x}) = p(\boldsymbol{x}) - p'(\boldsymbol{x}) \approx \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(\boldsymbol{x})$$
  - $$L^2(p,p') = \int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\Big) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \approx \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha}$$
    $$\boldsymbol{G} = \int \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x}$$
    $$\widehat{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{i'=1}^{n'} \boldsymbol{\phi}(\boldsymbol{x}'_{i'})$$

Bias Reduction

- Consider their linear combination:

$$\kappa \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} + (1-\kappa)\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{G}\widehat{\boldsymbol{\alpha}} \qquad \kappa \in \mathbb{R}$$

  - For small λ ,

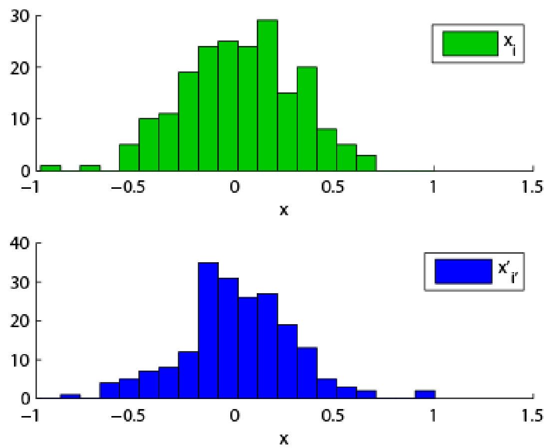
$$\kappa \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} + (1-\kappa)\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{G}\widehat{\boldsymbol{\alpha}}$$

$$= \widehat{\boldsymbol{h}}^\top \boldsymbol{G}^{-1}\widehat{\boldsymbol{h}} - \lambda(2-\kappa)\widehat{\boldsymbol{h}}^\top \boldsymbol{G}^{-2}\widehat{\boldsymbol{h}} + o_p(\lambda)$$

  - $\kappa = 2$ removes the regularization-induced bias:

$$\widehat{L}^2(\mathcal{X}, \mathcal{X}') = 2\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{G}\widehat{\boldsymbol{\alpha}}$$

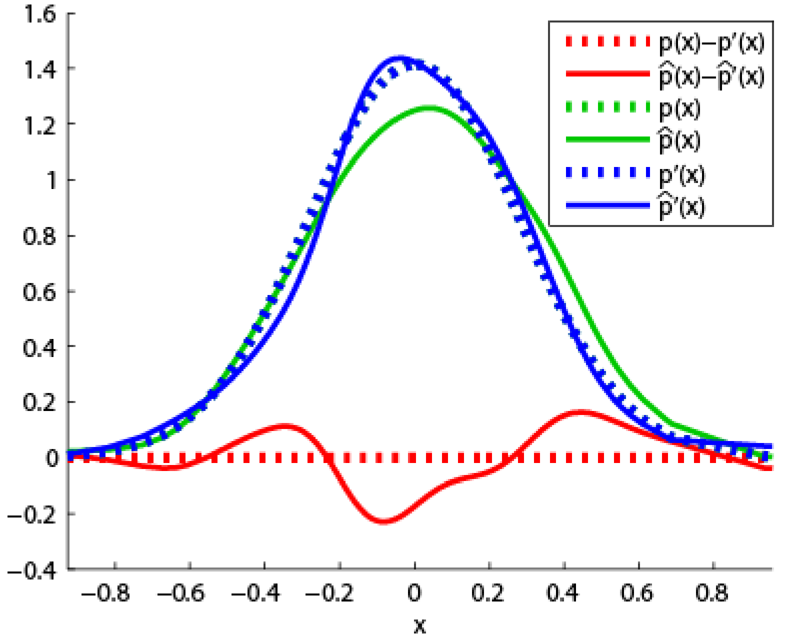
Density-Difference Estimation (1)

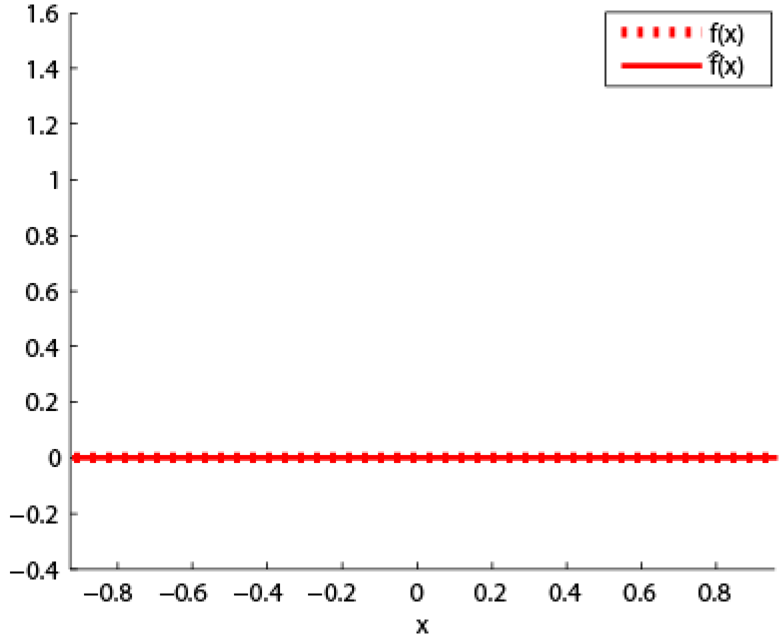


- $p(x) = p'(x) = N(x; 0, (4\pi)^{-1})$

$$n = n' = 200$$

Difference of kernel Density estimators (KDE)	Least-squares density -difference estimation (LSDD)

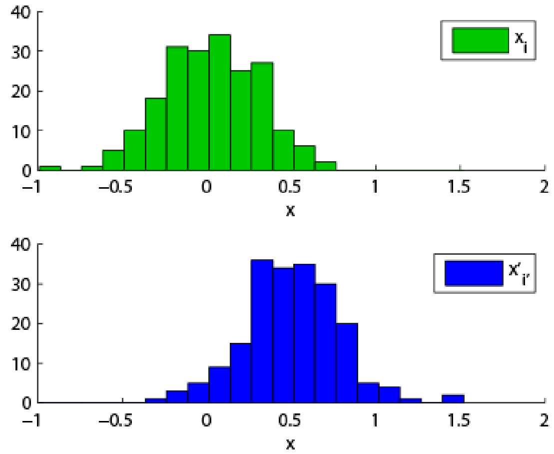




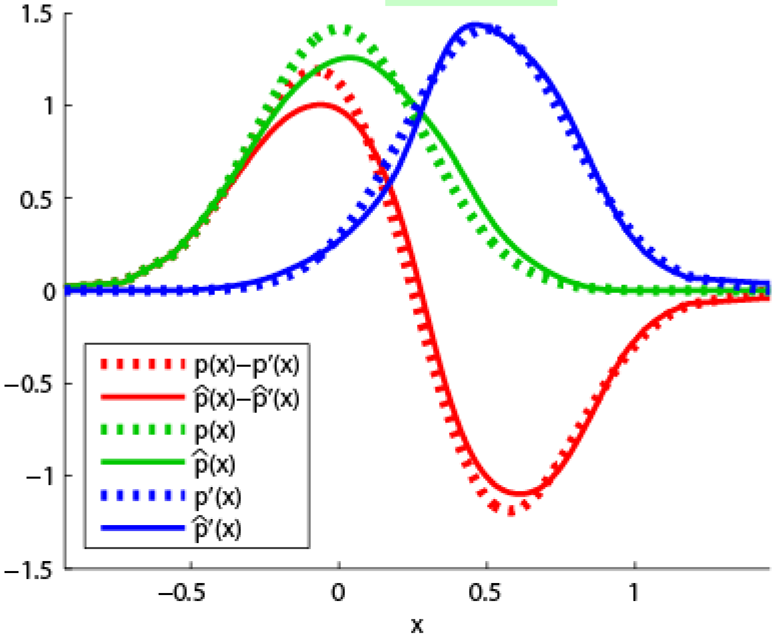
Density-Difference Estimation (2)

- $p(x) = N(x; 0, (4\pi)^{-1})$
- $p'(x) = N(x; 0.5, (4\pi)^{-1})$

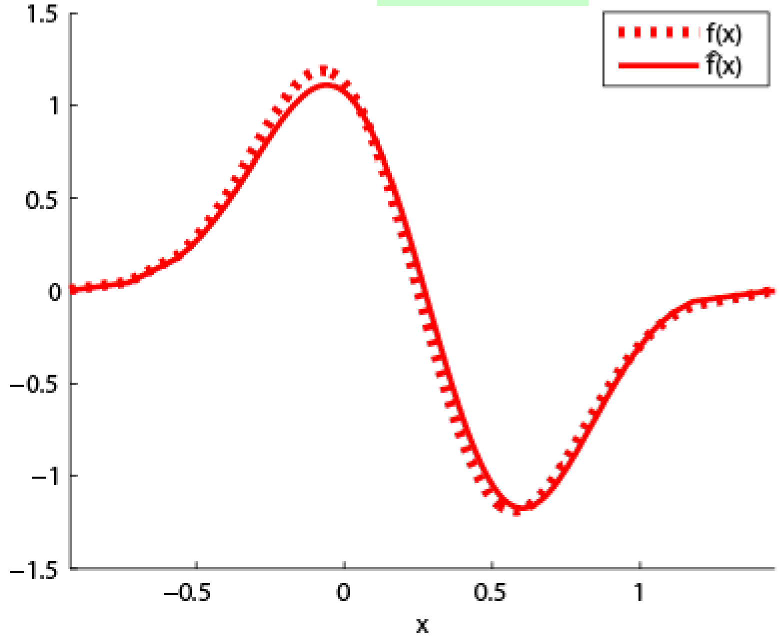
$$n = n' = 200$$



KDE

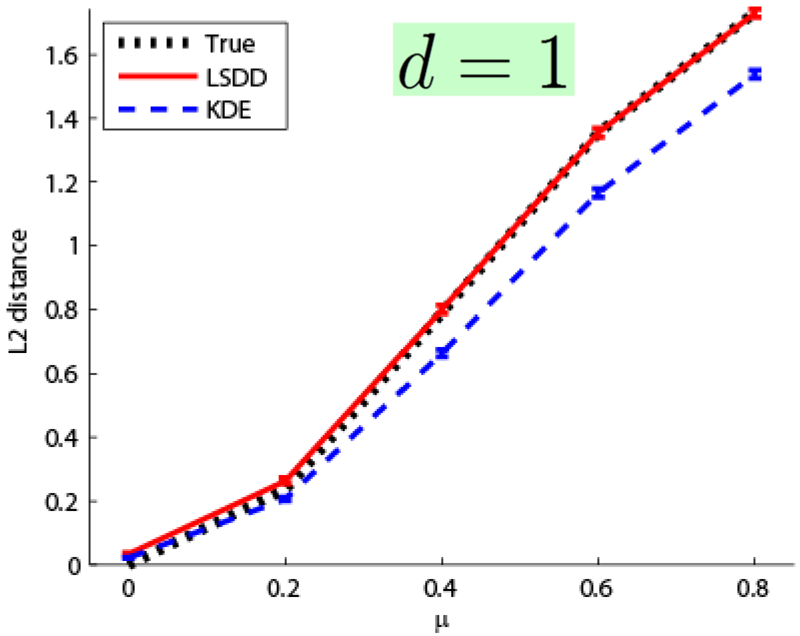


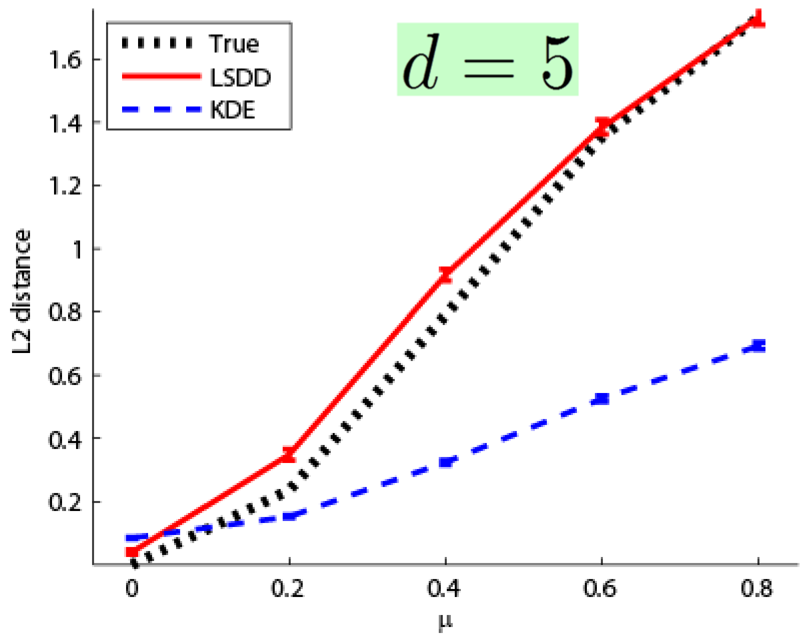
LSDD



L^2-Distance Approximation

- $p(\boldsymbol{x}) = N(\boldsymbol{x}; (\mu, 0, \ldots, 0)^\top, (4\pi)^{-1}\boldsymbol{I}_d)$ $\quad n = n' = 100$
- $p'(\boldsymbol{x}) = N(\boldsymbol{x}; (0, 0, \ldots, 0)^\top, (4\pi)^{-1}\boldsymbol{I}_d)$



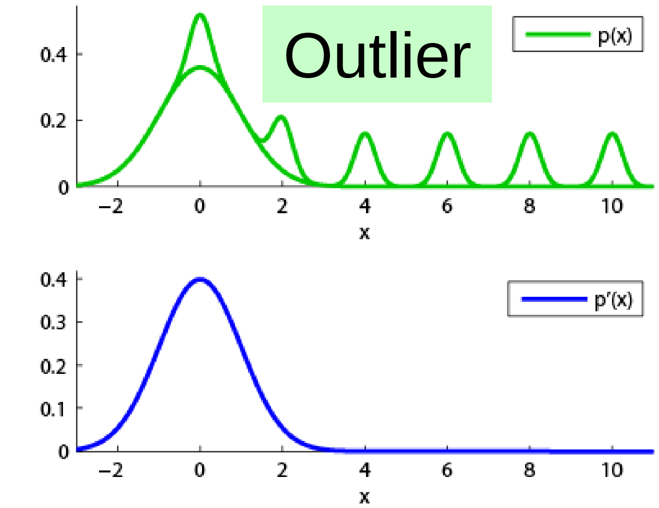


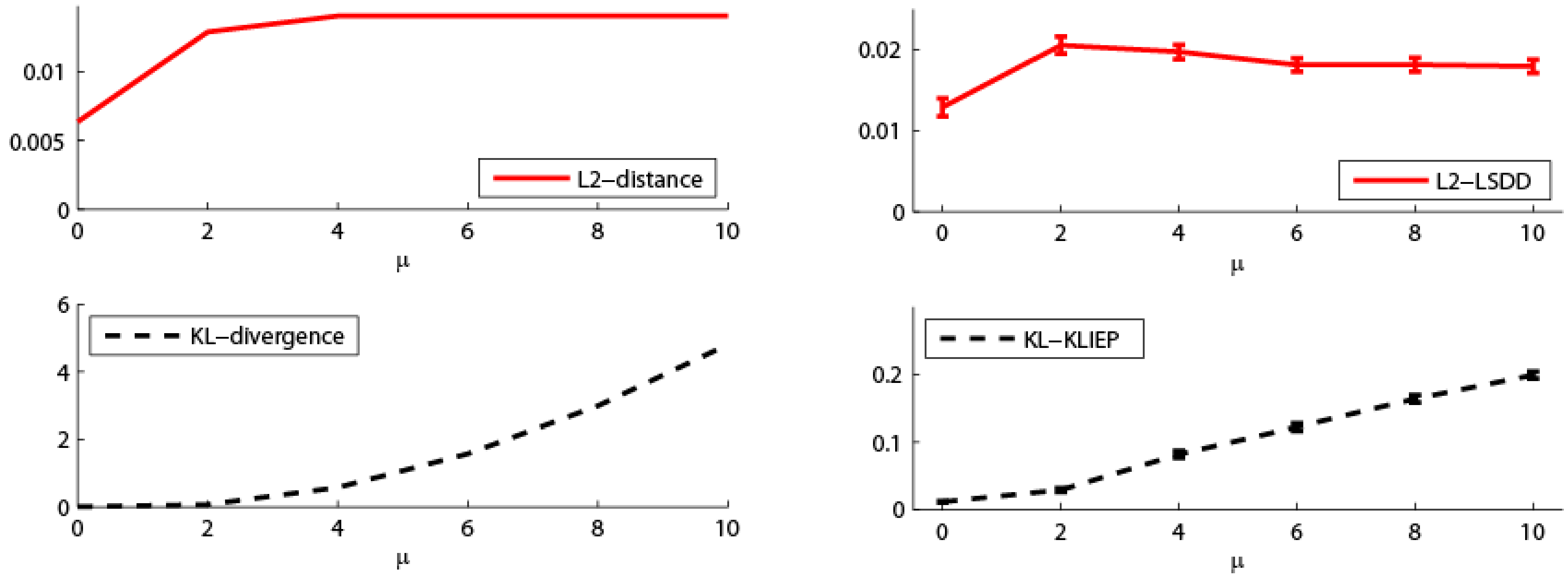
- KDE significantly under-estimates.
- LSDD slightly over-estimates.

L^2-Distance vs. KL-Divergence

$$L^2(p,p') = \int \left(p(\boldsymbol{x}) - p'(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}$$

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

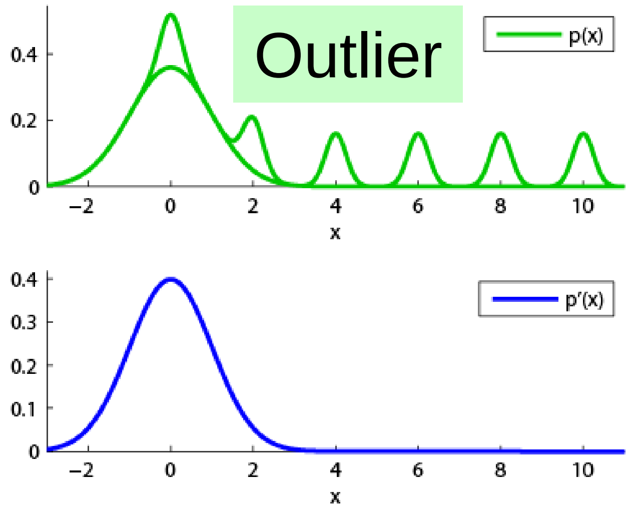


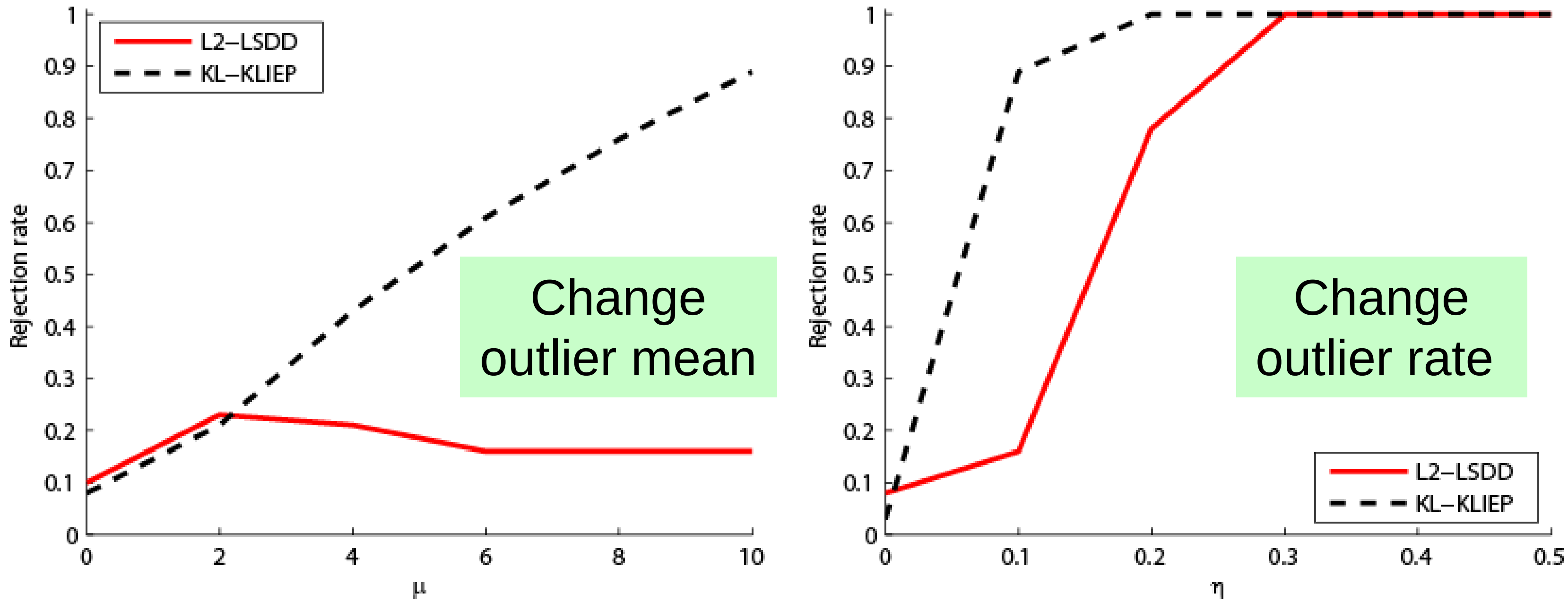


L^2-distance is less sensitive to outliers.

Robust Two-Sample Test

- **Two-sample test:** Are two distributions the same?
  - Null: Two are the same
  - Alternative: Two are different

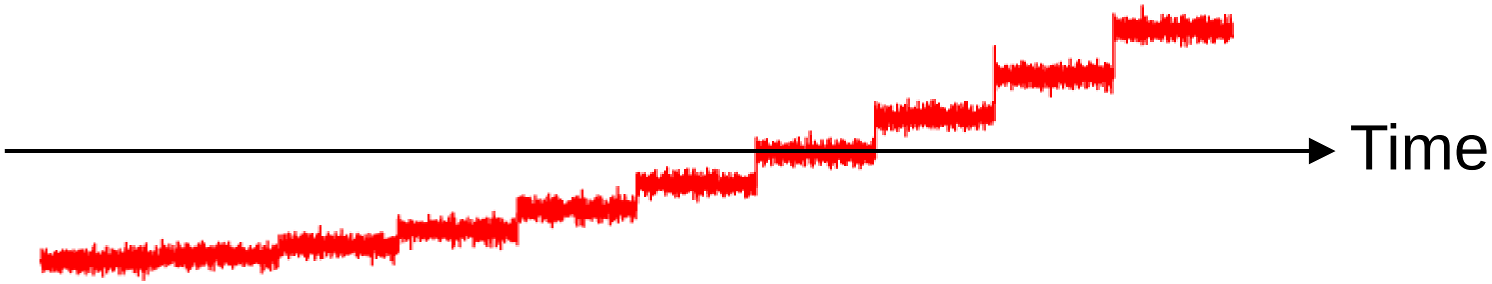




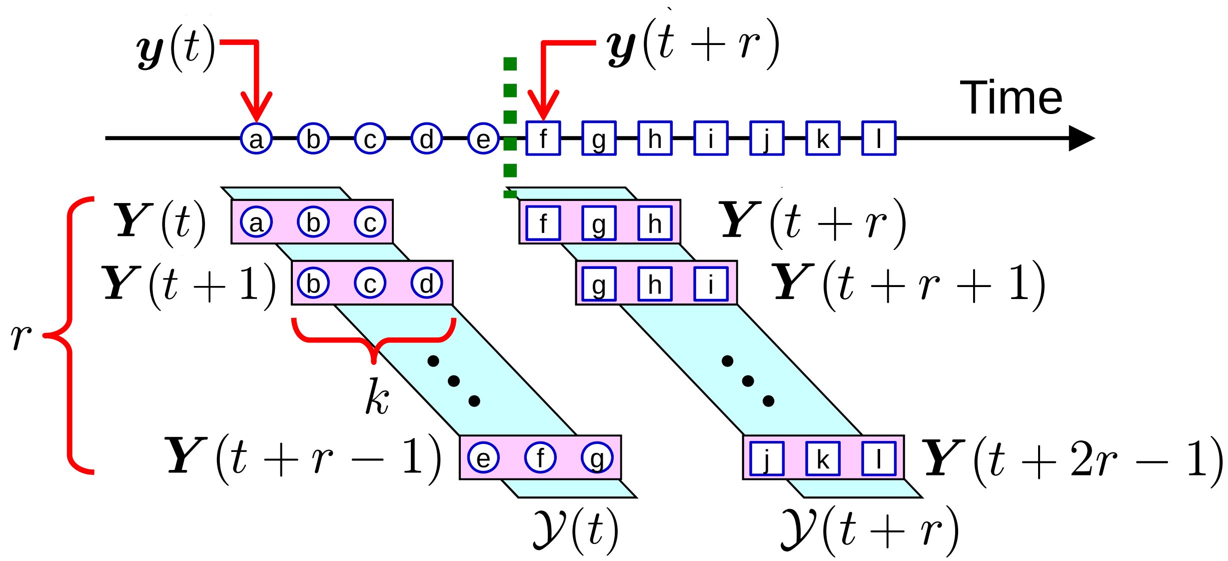
L^2-based test is more robust against outliers.

Unsupervised Change Detection

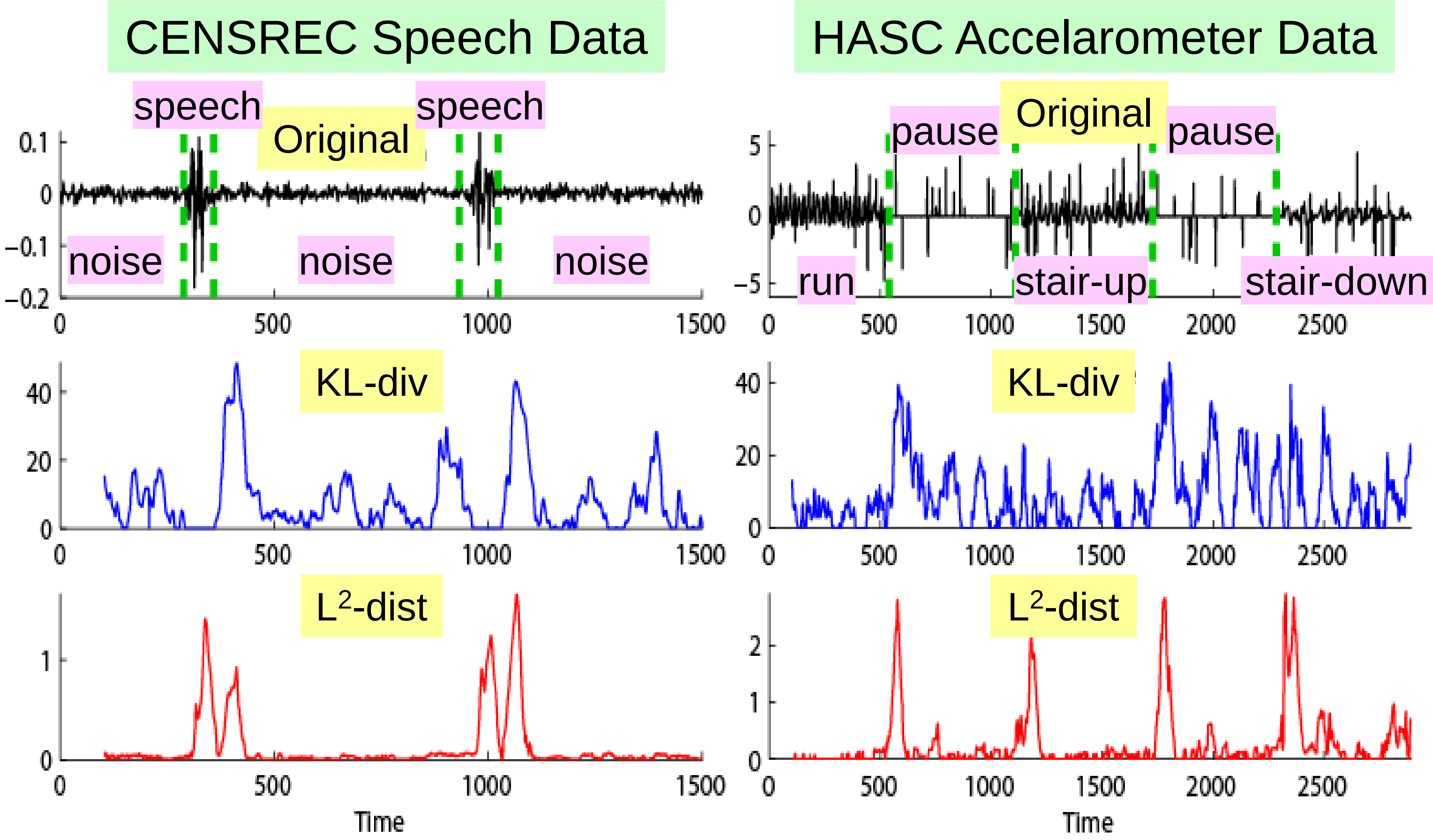
- Identify change points in time-series:



- Use the distance between the distributions of sliding-windowed past and current data.



Results



L^2 is more robust against noise.

Mutual Information

$$\mathrm{MI} = \iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

- Mutual information is the KL-divergence from the joint density $p(\boldsymbol{x}, \boldsymbol{y})$ to the product of marginal densities $p(\boldsymbol{x})p(\boldsymbol{y})$.
- Independence can be measured:
  - $\mathrm{MI} \geq 0$
  - $\mathrm{MI} = 0$ $\Longleftrightarrow$ $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent

Mutual Information Approximation

$$\mathrm{MI} = \iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

- Estimation of density ratio $r(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$

  from
  $$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, \boldsymbol{y})$$
  $$\{(\boldsymbol{x}_i, \boldsymbol{y}_{i'})\}_{i,i'=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})p(\boldsymbol{y})$$

  gives an MI approximator:

  $$\widehat{\mathrm{MI}} = \frac{1}{n} \sum_{i=1}^{n} \log \widehat{r}(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

Variations of MI

- Squared-loss MI (Pearson divergence):

$$\mathrm{SMI} = \iint p(\boldsymbol{x})p(\boldsymbol{y}) \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

- Relative SMI:

$$\mathrm{rSMI} = \iint p_\beta(\boldsymbol{x},\boldsymbol{y}) \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p_\beta(\boldsymbol{x},\boldsymbol{y})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

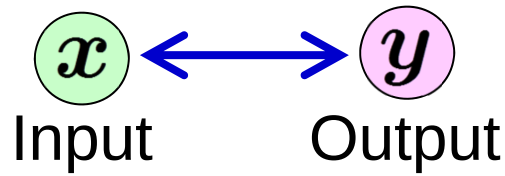
$0 \leq \beta < 1$ $\quad$ $p_\beta(\boldsymbol{x},\boldsymbol{y}) = \beta p(\boldsymbol{x},\boldsymbol{y}) + (1-\beta)p(\boldsymbol{x})p(\boldsymbol{y})$

- Quadratic MI:

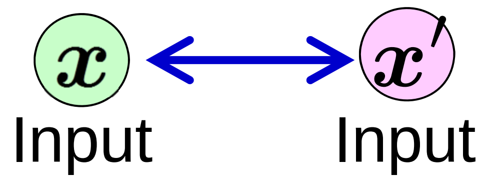
$$\mathrm{QMI} = \iint \Big(p(\boldsymbol{x},\boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y}) \Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

Usages of MI Approximator

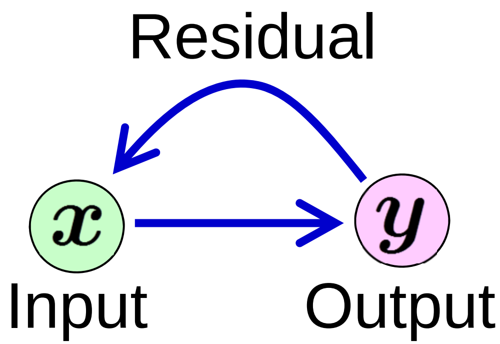
- MI between input and output:
  - Feature selection/extraction
  - Clustering



- MI between inputs:
  - Independent component analysis
  - Higher-order canonical correlation analysis
  - Object matching



- MI between input and residual:
  - Causal direction inference



Summary of Distributional Change Detection

- Compute a divergence between distributions:
  - Separate density estimation does not work well, because Vapnik's principle is violated.
  - Direct estimation of density ratio/difference seems more sensible.
- Don't simply use KL as a divergence measure just because it is popular.
  - Relative PE and L^2 could be more robust against outliers and computationally more efficient.
- MI can also be approximated in the same way.

A Little Break: Artist Agent

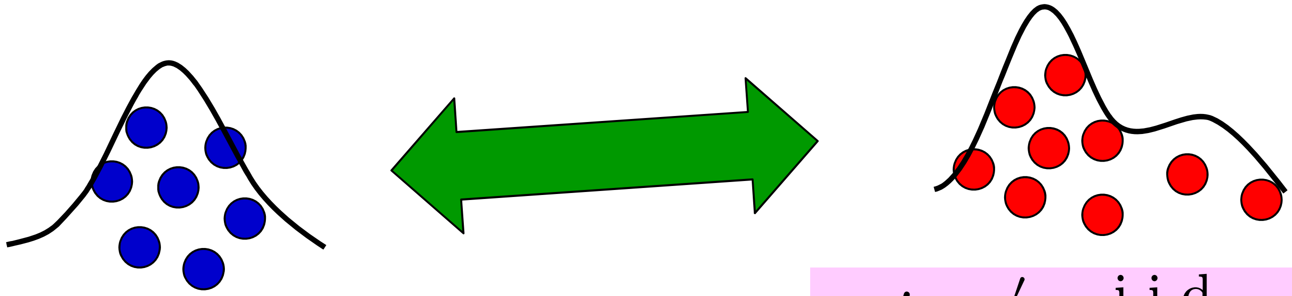
Ning *et al.* (ICML2012)

Brush movement learning by reinforcement learning.

Contents

1. Distributional change detection
2. Structural change detection
   - A) Density estimation approach
   - B) Density-ratio estimation approach

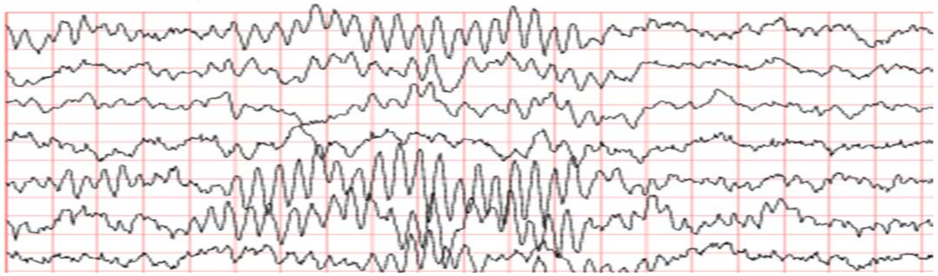
From Distributional Change to Structural Change



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

$\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$

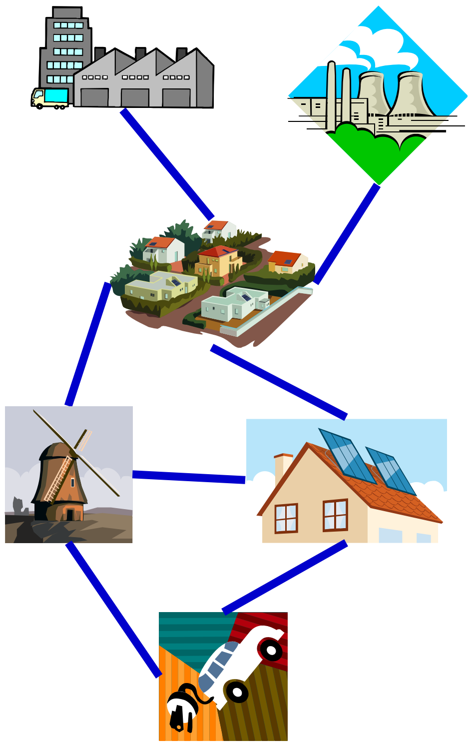
- Through distance estimation, distributional change can be detected.
- We investigate how distributions are changed through interaction between variables.

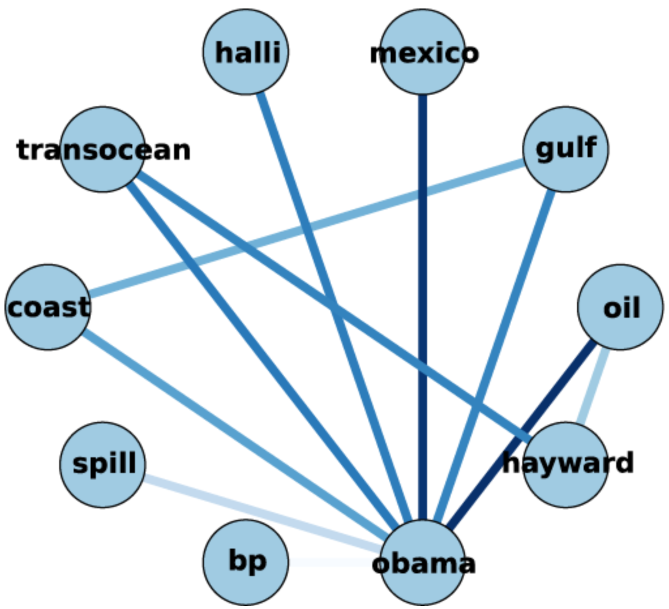


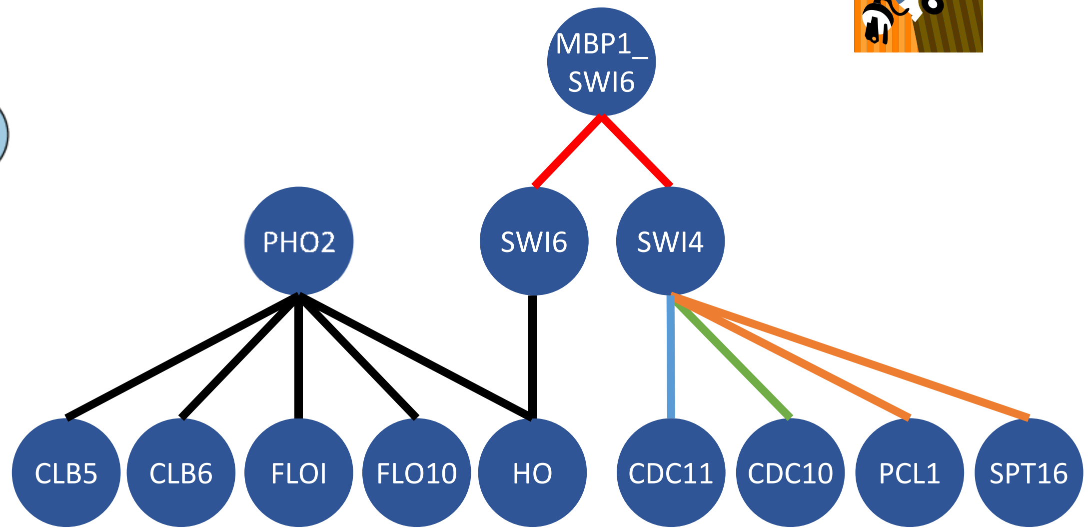
$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

Motivating Examples

- Word co-occurrence in Twitter
- Gene regulatory networks
- Fraud detection in smart grid







Contents

1. Distributional change detection
2. Structural change detection
   A) Density estimation approach
      I. Gauss models
      II. Non-Gauss models
   B) Density-ratio estimation approach

Gauss Model

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

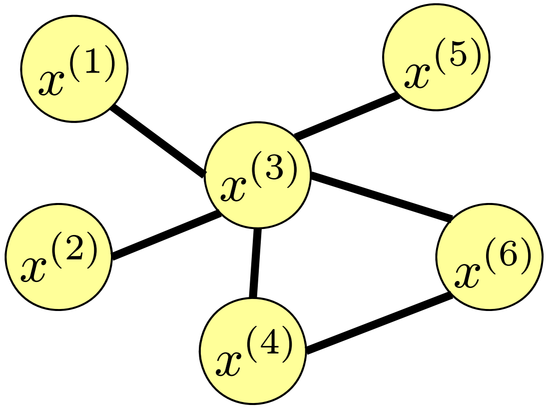
$\boldsymbol{\Theta}$: (sparse) inverse covariance matrix

- Conditional independence:

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

$$\Theta_{k,k'} = 0 \iff x^{(k)} \perp\!\!\!\perp x^{(k')} \mid \{x^{(\ell)}\}_{\ell \neq k,k'}$$

- Graphical representation:
  - Node: Each variable
  - Edge: Exists if $\Theta_{i,j} \neq 0$
  - Only connected variables affect!



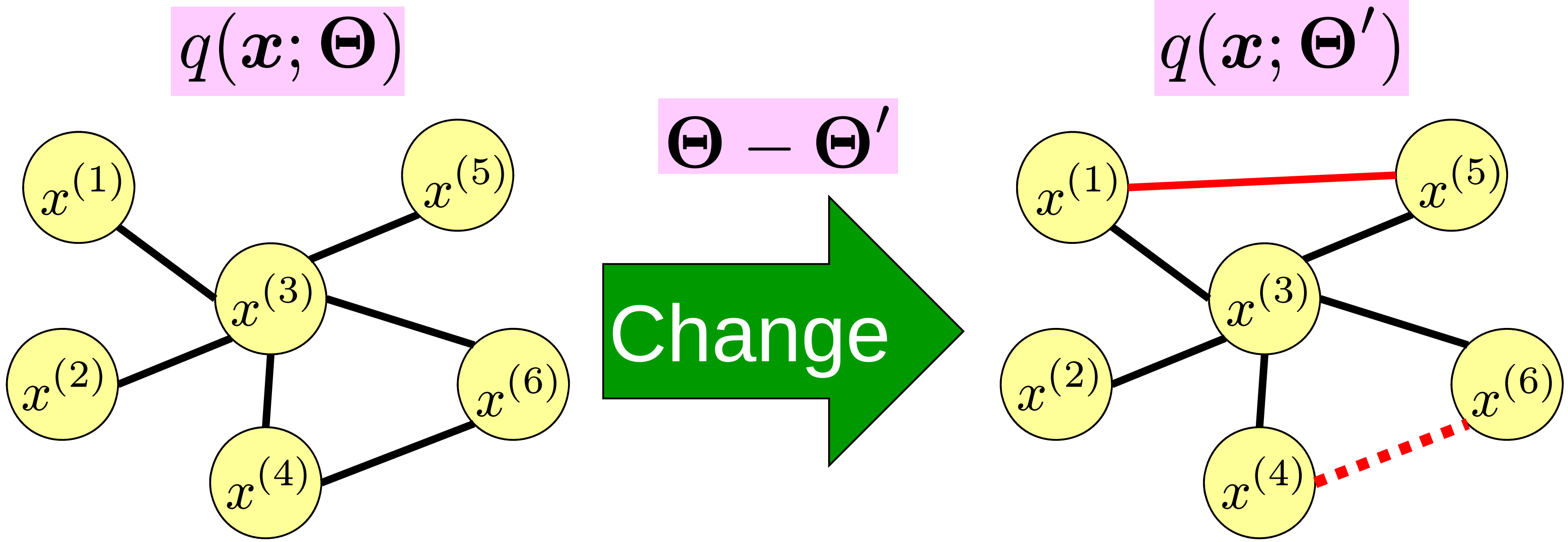
$$x^{(1)} \perp\!\!\!\perp x^{(2)} \mid x^{(3)}$$

Structural Change Detection with Gauss Models

- Use Gauss models for $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$:

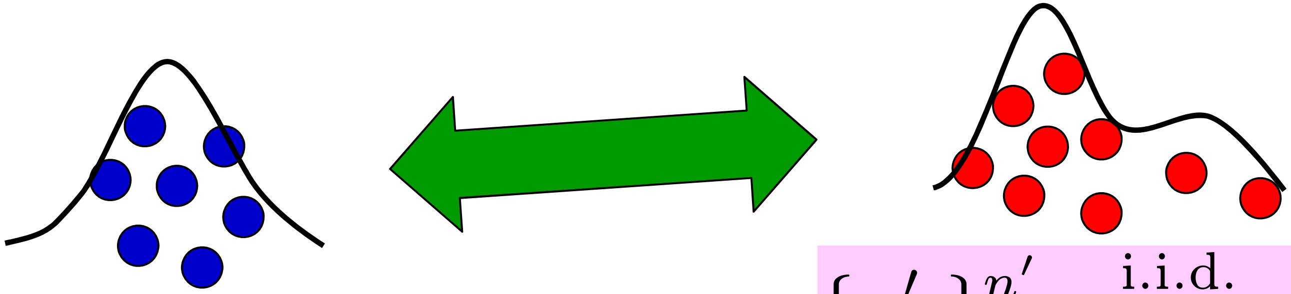
$$q(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta}\boldsymbol{x}\right) \qquad q(\boldsymbol{x};\boldsymbol{\Theta}')$$

- Detect sparse change in covariance structure:



Structural Change Detection by Graphical Lasso (Glasso)

Tibshirani (JRSS1996), Friedman *et al.* (Biostat2008)



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

- Sparse maximum likelihood estimation:

$$\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \log q(\boldsymbol{x}_i; \boldsymbol{\Theta}) - \lambda \|\boldsymbol{\Theta}\|_1 \qquad \max_{\boldsymbol{\Theta}'} \sum_{i'=1}^{n'} \log q(\boldsymbol{x}'_{i'}; \boldsymbol{\Theta}') - \lambda' \|\boldsymbol{\Theta}'\|_1$$

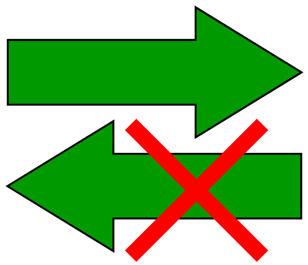
$$q(\boldsymbol{x}; \boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{x}\right) \qquad \lambda, \lambda' \geq 0$$

Structural Change Detection by Glasso

$$\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \log q(\boldsymbol{x}_i; \boldsymbol{\Theta}) - \lambda \|\boldsymbol{\Theta}\|_1 \qquad \max_{\boldsymbol{\Theta}'} \sum_{i'=1}^{n'} \log q(\boldsymbol{x}'_{i'}; \boldsymbol{\Theta}') - \lambda' \|\boldsymbol{\Theta}'\|_1$$

- ☺ Scalable to high-dimensional datasets.
- ☺ Statistical properties have been well studied.
- ☹ Does not work if true $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ are dense.
- ☹ Choice of λ and λ' is not straightforward.

Both $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ are sparse



Change $\boldsymbol{\Theta} - \boldsymbol{\Theta}'$ is sparse

Structural Change Detection by Fused Lasso (Flasso)

Tibshirani *et al.* (JRSS2005)
Zhang & Wang (UAI2010)

- Directly penalize the difference of parameters to be sparse:

$$\max_{\boldsymbol{\Theta},\boldsymbol{\Theta}'} \sum_{i=1}^{n} \log q(\boldsymbol{x}_i;\boldsymbol{\Theta}) + \sum_{i'=1}^{n'} \log q(\boldsymbol{x}'_{i'};\boldsymbol{\Theta}') - \gamma\|\boldsymbol{\Theta}-\boldsymbol{\Theta}'\|_1$$

$$\gamma \geq 0$$

☺ Scalable to high-dimensional datasets.

☺ Work well even if true $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ are dense.

Contents

1. Distributional change detection
2. Structural change detection
   - A) Density estimation approach
     - I. Gauss models
     - II. Non-Gauss models
   - B) Density-ratio estimation approach

Correlation and Dependence

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

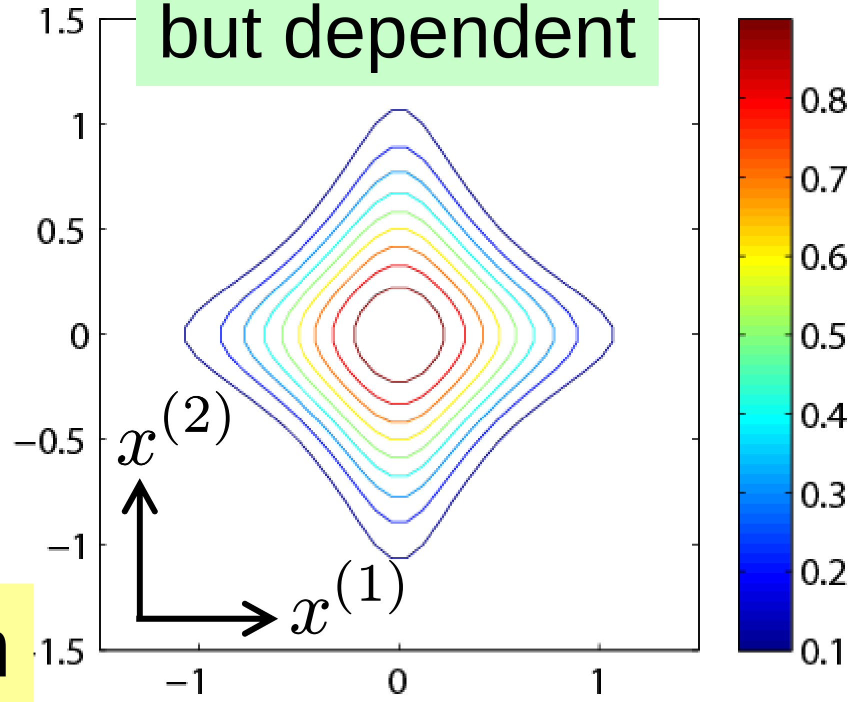
$\boldsymbol{\Theta}$: (sparse) inverse covariance matrix

- Gauss models cannot capture higher-order correlations.
- No correlation does not imply independence.

Independence $\Rightarrow$ No correlation

No correlation $\not\Rightarrow$ Independence

No correlation but dependent



Nonparanormal Models

Liu *et al.* (JMLR2009)

- Gaussian after element-wise transformation:

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

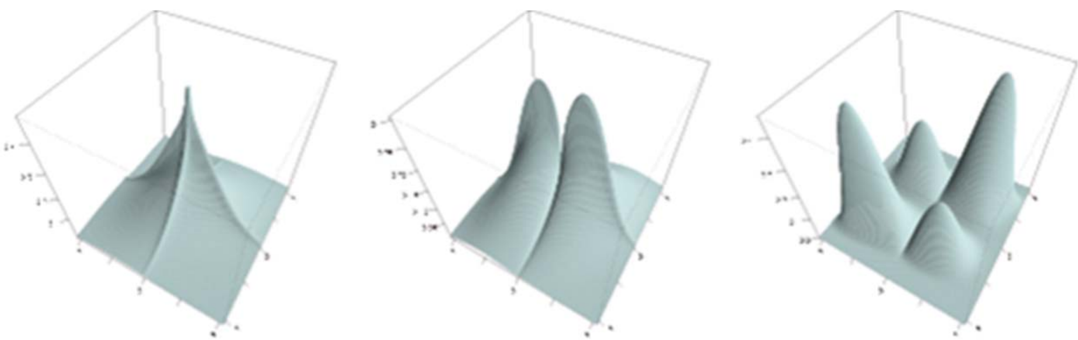
$$\boldsymbol{f}(\boldsymbol{x}) = (f_1(x^{(1)}), \ldots, f_d(x^{(d)}))^\top$$

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

f_k : Monotone and differentiable function

☺ More flexible than ordinary Gauss models.

☹ Still restrictive in representation



Pairwise Markov Networks

$$q(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\overline{q}(\boldsymbol{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \qquad \overline{q}(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{k \geq k'} \boldsymbol{\theta}_{k,k'}^\top \boldsymbol{f}(x^{(k)}, x^{(k')})\right)$$

$\boldsymbol{f}(x, x')$: feature vector

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_{1,1}^\top, \ldots, \boldsymbol{\theta}_{d,d}^\top)^\top$$

- Gaussian: $\boldsymbol{f}(x, x') = xx'$
- Nonparanormal: $\boldsymbol{f}(x, x') = f(x)f(x')$
- Polynomial: $\boldsymbol{f}(x, x') = [x^t, x^{t-1}x', \ldots, x, x', 1]^\top$

☺ High representation capability.

☹ Normalization $Z(\boldsymbol{\theta}) = \int \overline{q}(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}$ is intractable.

Importance Sampling

1. Draw pseudo-samples from a proposal density:

$$\{\boldsymbol{x}''_{i''}\}_{i''=1}^{n''} \overset{\text{i.i.d.}}{\sim} p''(\boldsymbol{x})$$

(e.g., Gaussian)

2. Approximate the integration by importance-weighted sample average:

$$Z(\boldsymbol{\theta}) = \int \overline{q}(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = \int \frac{\bar{q}(\boldsymbol{x};\boldsymbol{\theta})}{p''(\boldsymbol{x})} p''(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\approx \frac{1}{n''}\sum_{i''=1}^{n''} \frac{\bar{q}(\boldsymbol{x}''_{i''};\boldsymbol{\theta})}{p(\boldsymbol{x}''_{i''})} \overset{n''\to\infty}{\longrightarrow} \int \overline{q}(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}$$

☺ Law of large numbers guarantees consistency.

☹ Unstable due to large variance.

Score Matching

Hyvärinen (JMLR2005)

- Learn unnormalized density model $\overline{q}(\boldsymbol{x};\boldsymbol{\theta})$ by least-squares matching of score functions:

$$\min_{\boldsymbol{\theta}} \int p(\boldsymbol{x}) \|\boldsymbol{\psi}(\boldsymbol{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\|^2 \mathrm{d}\boldsymbol{x}$$

$$\boldsymbol{\psi}(\boldsymbol{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} \log \overline{q}(\boldsymbol{x};\boldsymbol{\theta}) \qquad \nabla_{\boldsymbol{x}} = (\partial_{x^{(1)}}, \ldots, \partial_{x^{(d)}})^\top$$

- Empirical version (use integration-by-parts):

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} S(\boldsymbol{x}_i;\boldsymbol{\theta}) \qquad S(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{d} \left(\psi_k(\boldsymbol{x};\boldsymbol{\theta})^2 + 2\partial_{x^{(k)}} \psi_k(\boldsymbol{x};\boldsymbol{\theta}) \right)$$

$$\int \psi_k(\boldsymbol{x};\boldsymbol{\theta}) \partial_{x^{(k)}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int \partial_{x^{(k)}} \psi_k(\boldsymbol{x};\boldsymbol{\theta}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

☺ No normalization is needed.

Contents

1. Distributional change detection
2. Structural change detection
   A) Density estimation approach
   B) Density-ratio estimation approach

Avoiding Density Estimation

- Fused lasso + Score matching:

$\gamma \geq 0$

$$\max_{\boldsymbol{\Theta},\boldsymbol{\Theta}'} \sum_{i=1}^{n} S(\boldsymbol{x}_i;\boldsymbol{\theta}) + \sum_{i'=1}^{n'} S(\boldsymbol{x}'_{i'};\boldsymbol{\theta}') - \gamma\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_1$$

☺ Work well even if true $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ are dense.

☺ Higher-order correlations can be captured.

☹ Still need explicit modeling of $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$.

- Vapnik's principle:

Don't solve a more general problem

Direct Change Modeling in Markov Networks

Liu *et al.* (ECML2013, NeCo2014)

- Without separately modeling $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$, we directly model the density ratio $p(\boldsymbol{x})/p'(\boldsymbol{x})$:

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \approx \frac{q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta}')} \propto \exp\left(\sum_{k \geq k'} (\boldsymbol{\theta}_{k,k'} - \boldsymbol{\theta}'_{k,k'})^\top \boldsymbol{f}(x^{(k)}, x^{(k')})\right)$$

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

- Individual parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$ are not necessary, but their difference $\boldsymbol{\alpha} = \boldsymbol{\theta} - \boldsymbol{\theta}'$ is enough.

Ratio of Markov Network Models

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \frac{1}{N(\boldsymbol{\alpha})} \exp\left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^\top \boldsymbol{f}(x^{(k)}, x^{(k')}) \right)$$

- Normalization:

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{1,1}^\top, \ldots, \boldsymbol{\alpha}_{d,d}^\top)^\top$$

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \Longrightarrow \int p'(\boldsymbol{x}) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1$$

☺ Simple sample averaging is consistent:

$$N(\boldsymbol{\alpha}) = \int p'(\boldsymbol{x}) \exp\left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^\top \boldsymbol{f}(x^{(k)}, x^{(k')}) \right) \mathrm{d}\boldsymbol{x}$$

$$\approx \frac{1}{n'} \sum_{i'=1}^{n'} \exp\left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^\top \boldsymbol{f}({x'}_{i'}^{(k)}, {x'}_{i'}^{(k')}) \right)$$

Sparse Density-Ratio Estimation

Sugiyama *et al.* (NIPS2007, AISM2008)

- Density-ratio matching under KL-divergence:

$$\min_{\boldsymbol{\alpha}} \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x}) r_{\boldsymbol{\alpha}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) \approx \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$$

- Sample approximation gives

$$\min_{\boldsymbol{\alpha}} \log \frac{1}{n'} \sum_{i'=1}^{n'} \exp\left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^{\top} \boldsymbol{f}(x_{i'}'^{(k)}, x_{i'}'^{(k')}) \right) - \frac{1}{n} \sum_{i=1}^{n} \sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^{\top} \boldsymbol{f}(x_i^{(k)}, x_i^{(k')})$$

  - Tractable for any feature $\boldsymbol{f}(x^{(k)}, x^{(k')})$.

- Add a smoothing regularizer: $+\eta \|\boldsymbol{\alpha}\|^2$
- Add a group-sparsity regularizer: $+\gamma \sum_{k \geq k'} \|\boldsymbol{\alpha}_{k,k'}\|$

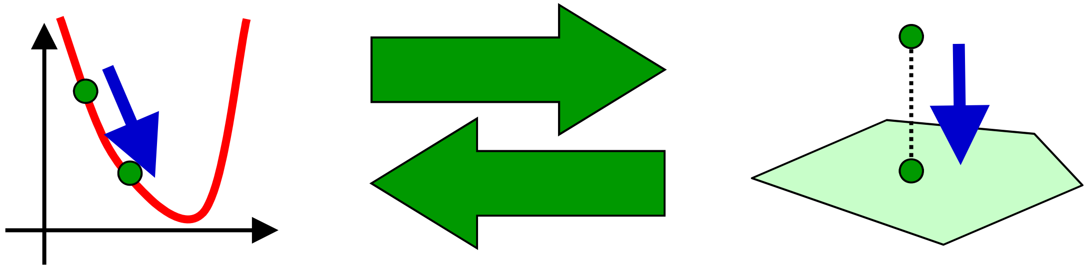
Primal Optimization

$$\min_{\boldsymbol{\alpha}} \log \frac{1}{n'} \sum_{i'=1}^{n'} \exp\left(\sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^{\top} \boldsymbol{f}(x_{i'}^{\prime(k)}, x_{i'}^{\prime(k')}) \right)$$

$$-\frac{1}{n} \sum_{i=1}^{n} \sum_{k \geq k'} \boldsymbol{\alpha}_{k,k'}^{\top} \boldsymbol{f}(x_{i}^{(k)}, x_{i}^{(k')}) + \eta \|\boldsymbol{\alpha}\|^2$$

$$\text{subject to} \quad \sum_{k \geq k'} \|\boldsymbol{\alpha}_{k,k'}\| \leq C_{\gamma}$$

- Simple gradient-projection gives the global solution.
- Efficient when more samples than parameters.



Dual Optimization

$$\min_{\boldsymbol{\beta}} \sum_{i'=1}^{n'} \beta_{i'}^{\top} \log \beta_{i'} + \frac{1}{2\eta} \sum_{k \geq k'} \max(0, \|\boldsymbol{m}_{k,k'}\| - \gamma)^2$$

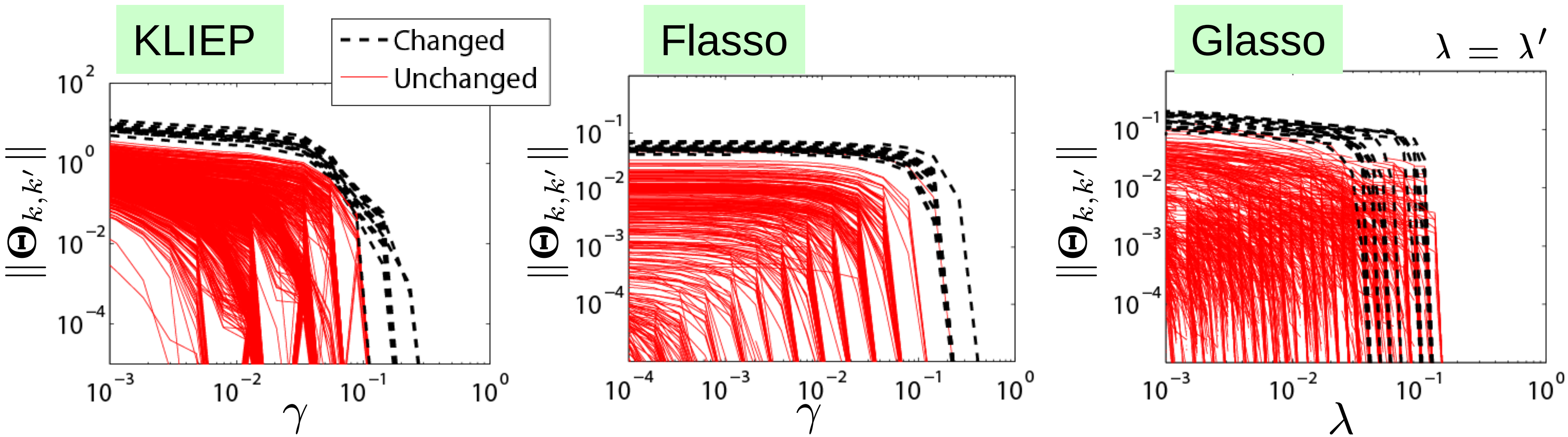
$$\text{subject to } \beta_1, \ldots, \beta_{n'} \geq 0, \sum_{i'=1}^{n'} \beta_{i'} = 1$$

$$\boldsymbol{m}_{k,k'} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i^{(k)}, \boldsymbol{x}_i^{(k')}) - \frac{1}{n'} \sum_{i'=1}^{n'} \beta_{i'} \boldsymbol{f}(\boldsymbol{x}_{i'}'^{(k)}, \boldsymbol{x}_i'^{(k')})$$

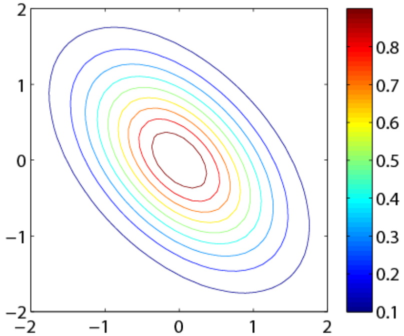
$$\Rightarrow \quad \boldsymbol{\alpha}_{k,k'} = \max(0, \|\boldsymbol{m}_{k,k'}\| - \gamma) \frac{\boldsymbol{m}_{k,k'}}{\eta \|\boldsymbol{m}_{k,k'}\|}$$

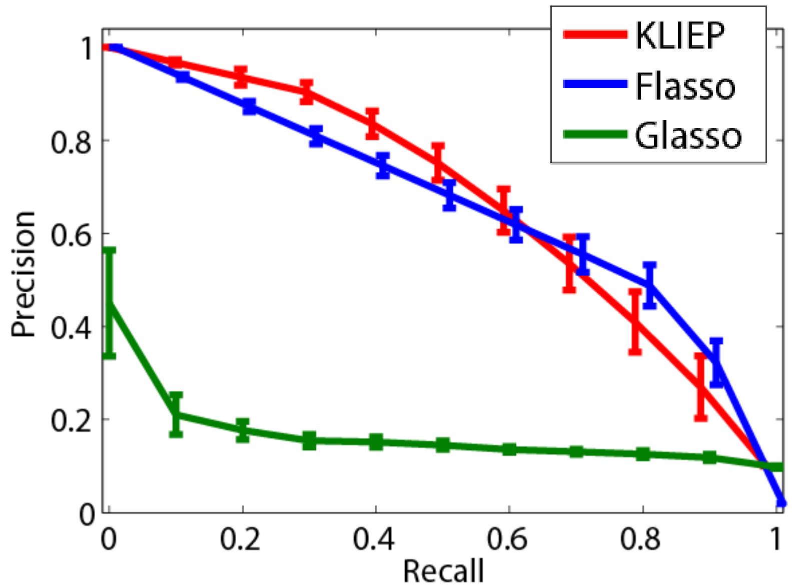
- Simple gradient-projection gives the global solution.
- Efficient when more parameters than samples.

Gaussian Data
(d=40, n=n'=100, Change in 15 Edges)

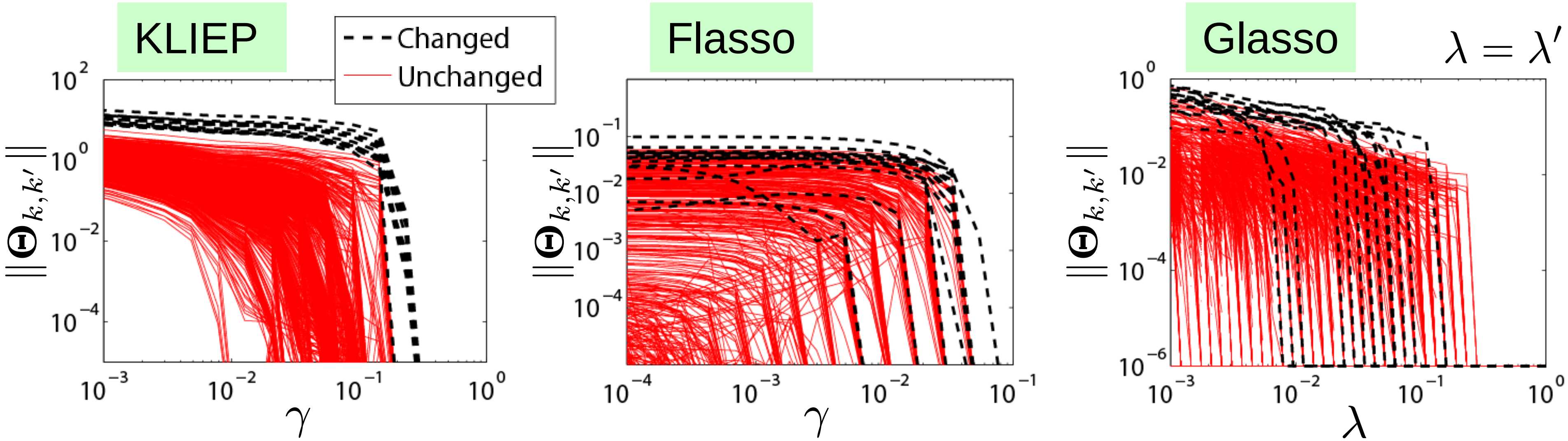


- All use the Gaussian model.
- KLIEP and Flasso work well.



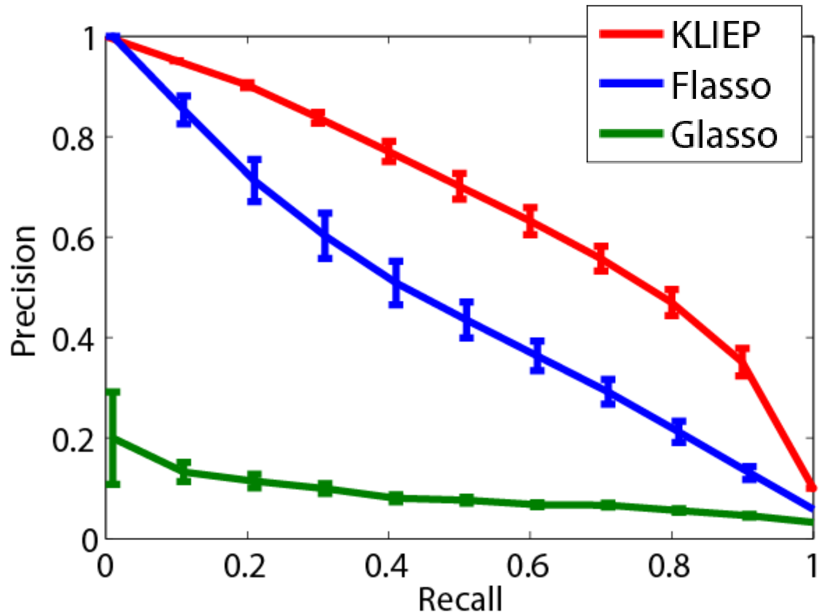


Gaussian Data (d=40, n=n'=50, Change in 15 Edges)

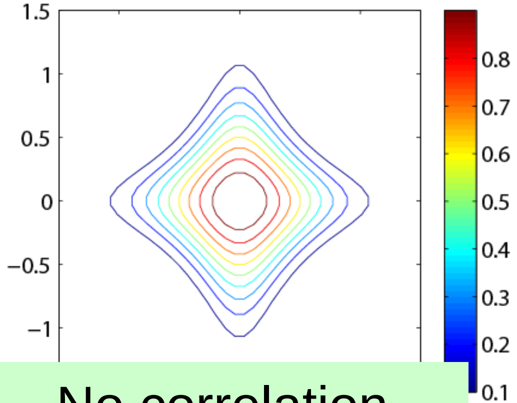


- KLIEP works well even with small samples.

$$\boldsymbol{\alpha} = \boldsymbol{\theta} - \boldsymbol{\theta}'$$



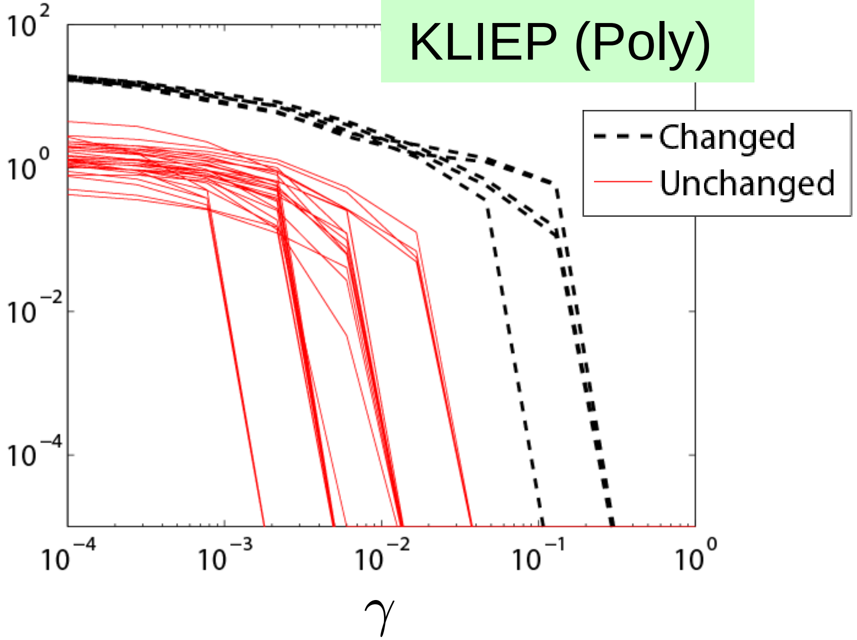
Non-Gaussian Data (d=9, n=n'=5000, Change in 7 Edges)

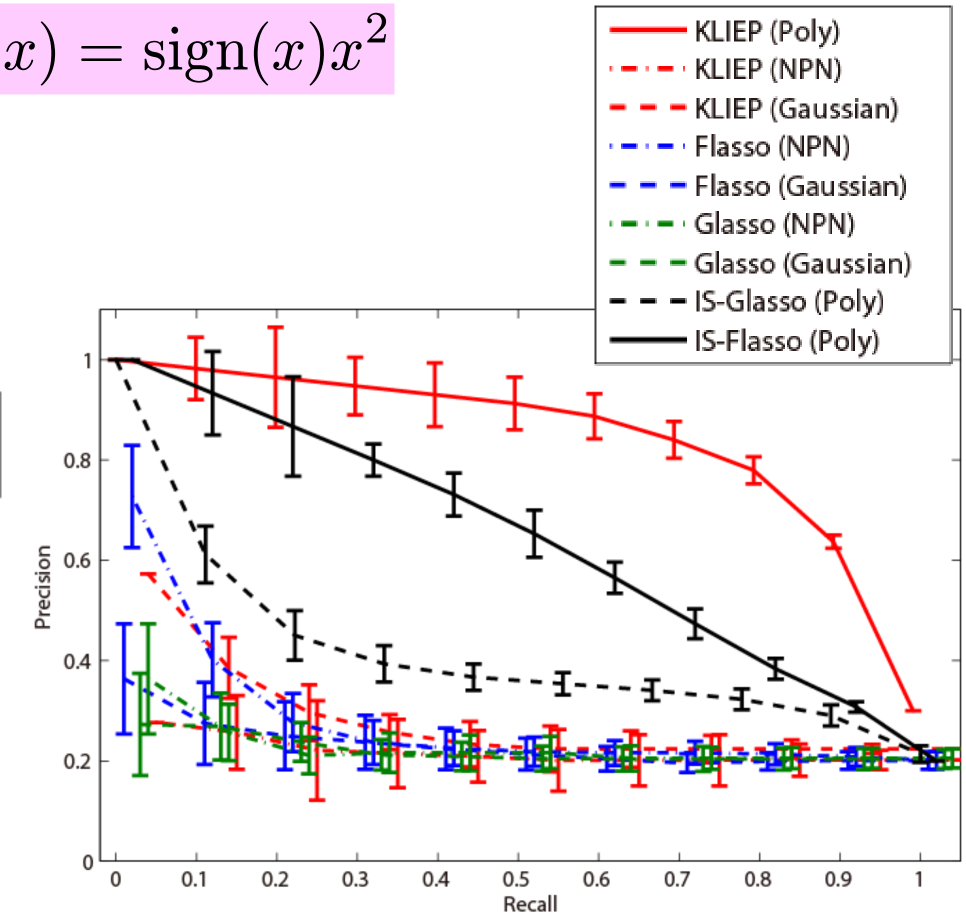


No correlation, no nonparanormal

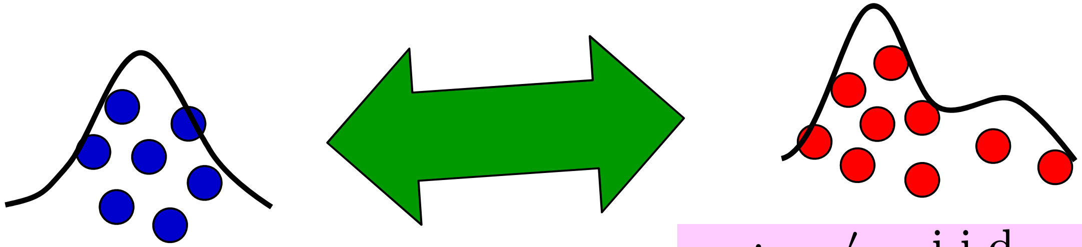
- KLIEP (Poly) works well.
  - Poly: $\boldsymbol{f}(x,x') = [x^t, x^{t-1}x', \ldots, x, x', 1]^\top$
  - NPN: $f(x) = \mathrm{sign}(x)x^2$

KLIEP (Poly)





Take-Home Messages



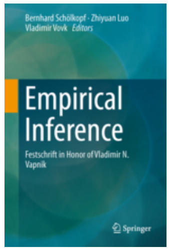
$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$

$\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$

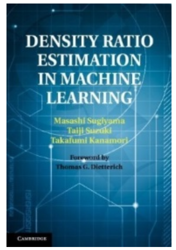
Don't solve a more general problem

- Directly learn the change:
  - Flexible and robust distributional change detection by direct density-ratio/density-difference estimation
  - Interpretable and tractable structural change detection by group-sparse density-ratio estimation
- Software: http://sugiyama-www.cs.titech.ac.jp/~sugi/software/

Schölkopf *et al.* (eds.), Empirical Inference, Festschrift in Honor of Vladimir N. Vapnik, Springer, 2013



Sugiyama *et al.*, Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012



PE-Divergence Approximation

Kanamori *et al.* (NIPS2008, JMLR2009)

$$\mathrm{PE}(p\|p') = \int p'(\boldsymbol{x})\Big(r(\boldsymbol{x}) - 1\Big)^2 \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x})r(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - 1$$

- Directly approximate the density ratio by least-squares:

$$\widehat{r} = \mathop{\mathrm{argmin}}_{\widetilde{r}} \int p'(\boldsymbol{x})\Big(\widetilde{r}(\boldsymbol{x}) - r(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}$$

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$$

$$= \mathop{\mathrm{argmin}}_{\widetilde{r}} \int p'(\boldsymbol{x})\Big(\widetilde{r}(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x} - 2\int p(\boldsymbol{x})r(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\Rightarrow \quad \mathrm{PE}(p\|p') \approx \int p(\boldsymbol{x})\widehat{r}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - 1$$

- Expectation is approximated by empirical average.

Contents

1. Distributional change detection
   - A) Problem setup and motivating examples
   - B) Distance approximation
     - I. Kullback-Leibler divergence
     - II. Pearson divergence
     - III. Relative Pearson divergence
     - IV. L^2-distance
2. Structural change detection

rPE-Divergence Approximation

Yamada *et al.* (NIPS2011, NeCo2013)

$$\mathrm{rPE}(p\|p') = \int p_\beta(\boldsymbol{x})\left(\frac{p(\boldsymbol{x})}{p_\beta(\boldsymbol{x})} - 1\right)^2 \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x})\frac{p(\boldsymbol{x})}{p_\beta(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - 1$$

$$p_\beta(\boldsymbol{x}) = \beta p(\boldsymbol{x}) + (1-\beta)p'(\boldsymbol{x}) \qquad 0 \le \beta < 1$$

- Directly approximate the relative density ratio by LS:

$$\widehat{r}_\beta = \underset{\widetilde{r}}{\mathrm{argmin}} \int p_\beta(\boldsymbol{x})\left(\widetilde{r}(\boldsymbol{x}) - \frac{p(\boldsymbol{x})}{p_\beta(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x}$$

$$= \underset{\widetilde{r}}{\mathrm{argmin}} \int p_\beta(\boldsymbol{x})\left(\widetilde{r}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} - 2\int p(\boldsymbol{x})\widetilde{r}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

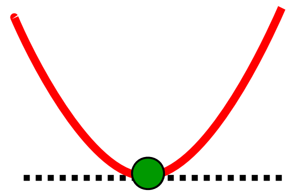
$$\Longrightarrow \quad \mathrm{rPE}(p\|p') \approx \int p(\boldsymbol{x})\widehat{r}_\beta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - 1$$

- Expectation is approximated by empirical average.

Solution for Linear Model

$$\widehat{\boldsymbol{\alpha}}_\beta = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\, \frac{\beta}{n'} \sum_{i'=1}^{n'} r_{\boldsymbol{\alpha}}(\boldsymbol{x}'_{i'})^2 + \frac{1-\beta}{n} \sum_{i=1}^{n} r_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)^2 - \frac{2}{n} \sum_{i=1}^{n} r_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)$$

$$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$



- (Regularized) solution is given analytically:

$$\widehat{\boldsymbol{\alpha}}_\beta = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}} \left[\boldsymbol{\alpha}^\top \widehat{\boldsymbol{G}}_\beta \boldsymbol{\alpha} - 2\widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right]$$

$$\widehat{\boldsymbol{h}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)$$

$$= (\widehat{\boldsymbol{G}}_\beta + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$$

$$\widehat{\boldsymbol{G}}_\beta = \frac{\beta}{n'} \sum_{i'=1}^{n'} \boldsymbol{\phi}(\boldsymbol{x}'_{i'}) \boldsymbol{\phi}(\boldsymbol{x}'_{i'})^\top + \frac{1-\beta}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i) \boldsymbol{\phi}(\boldsymbol{x}_i)^\top$$

- Resulting rPE-divergence approximator:

$$\mathrm{rPE}(p \| p') \approx \frac{1}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{\alpha}}_\beta^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - 1 = \widehat{\boldsymbol{h}}^\top (\widehat{\boldsymbol{G}}_\beta + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}} - 1$$