December 3-6, 2012
Lake Tahoe, Nevada, USA

Density-Difference Estimation

Masashi Sugiyama (Tokyo Institute of Technology, Japan), Takafumi Kanamori (Nagoya University, Japan), Taiji Suzuki (University of Tokyo, Japan)
Marthinus Christoffel du Plessis (Tokyo Institute of Technology, Japan), Song Liu (Tokyo Institute of Technology, Japan), Ichiro Takeuchi (Nagoya Institute of Technology, Japan)

1. This Work in A Nutshell

- Target problem: Estimate the difference between two densities.
- Approach: Avoid density estimation and directly estimate the difference in a single-shot process.
- Theory: Parametric and non-parametric optimality in terms of the approximation accuracy.
- Usage: L^2-distance approximation.
- Applications: Semi-supervised class-prior estimation and unsupervised change detection.

2.1 Target Problem & Motivations

- From two sets of i.i.d. samples

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

estimate the difference between two densities:

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

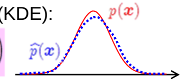
- Via density-difference estimation, we want to
  - Compare probability distributions.
  - Approximate the L^2-distance.

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

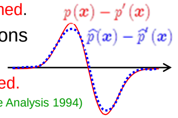
2.2 Naïve Approach

- 2-step method of first estimating two densities separately and then computing their difference.
- Kernel density estimation (KDE):

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



- A smoother function is obtained.
- Difference of smooth functions tends to be too smooth.
  - L^2-distance is under-estimated.



Cf. Anderson, Hall & Titterington (J. Multivariate Analysis 1994)

2.3 Vapnik's Principle

Vladimir N. Vapnik (1998)
Statistical Learning Theory, Wiley.

If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

3.1 Least-Squares Density-Difference (LSDD) Estimation

- Density-difference model:

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{n+n'} \theta_\ell \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_\ell\|^2}{2\sigma^2}\right)$$

$$(\boldsymbol{c}_1,\ldots,\boldsymbol{c}_n,\boldsymbol{c}_{n+1},\ldots,\boldsymbol{c}_{n+n'}) = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{x}'_1,\ldots,\boldsymbol{x}'_{n'})$$

- Least-squares estimation:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \qquad J(\boldsymbol{\theta}) = \int \left(g(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}$$

$$\boldsymbol{\theta} = (\theta_1,\ldots,\theta_{n+n'})^\top \qquad f(\boldsymbol{x}) = p(\boldsymbol{x}) - p'(\boldsymbol{x})$$

- Sample approximation:

$$J(\boldsymbol{\theta}) = \underbrace{\int g(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}}_{\text{Analytically computable}} - 2\underbrace{\int g(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}_{\text{Consistently estimable from samples}} + \underbrace{\int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}}_{\text{Safely ignorable}}$$

- $$\int g(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} = \sum_{\ell,\ell'=1}^{n+n'} \theta_\ell \theta_{\ell'} (\pi\sigma^2)^{d/2} \exp\left(-\frac{\|\boldsymbol{c}_\ell - \boldsymbol{c}_{\ell'}\|^2}{4\sigma^2}\right) \qquad \boldsymbol{x} \in \mathbb{R}^d$$
- $$\int g(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{i'=1}^{n'} g(\boldsymbol{x}'_{i'})$$

- Regularized training criterion:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left[\boldsymbol{\theta}^\top \boldsymbol{H}\boldsymbol{\theta} - 2\widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}\right]$$

$$H_{\ell,\ell'} = (\pi\sigma^2)^{d/2} \exp\left(-\frac{\|\boldsymbol{c}_\ell - \boldsymbol{c}_{\ell'}\|^2}{4\sigma^2}\right)$$

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

- Solution can be computed analytically:

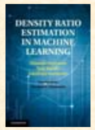
$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{H} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}$$

- Cross-validation is available for objectively tuning Gaussian width σ^2 and regularization parameter λ.
- MATLAB code is available:

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

Related Work: Density Ratio Estimation

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



Sugiyama, Suzuki & Kanamori (2012)
Density Ratio Estimation in Machine Learning
Cambridge University Press

- Density ratio is a versatile tool:
  - Importance sampling: $\sum_{i=1} \frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \mathrm{loss}(\boldsymbol{x}_i)$
  - Divergence estimation: $\int p'(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$
  - Mutual information estimation: $\iint p(\boldsymbol{x},\boldsymbol{y}) \log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$
  - Conditional probability estimation: $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$

3.2 Theoretical Analyses

- Parametric case: $g(\boldsymbol{x}) = \sum_{\ell=1}^{b} \theta_\ell \psi_\ell(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{\psi}(\boldsymbol{x})$

For $\frac{n}{n+n'} \xrightarrow{n,n' \to \infty} \eta \in [0,1]$, we have

$$\left(\sqrt{\frac{1}{n} + \frac{1}{n'}}\right)^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \rightsquigarrow N\left(\boldsymbol{0}, \boldsymbol{H}^{-1}((1-\eta)\boldsymbol{V}_p + \eta \boldsymbol{V}_{p'})\boldsymbol{H}^{-1}\right)$$

$\widehat{\boldsymbol{\theta}}$: LSDD estimator $\qquad \boldsymbol{V}_p = \int \left(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}_p\right)\left(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}_p\right)^\top p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$

$\boldsymbol{\theta}^*$: Optimal parameter $\qquad \boldsymbol{H} = \int \boldsymbol{\psi}(\boldsymbol{x})\boldsymbol{\psi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x} \quad \boldsymbol{\psi}_p = \int \boldsymbol{\psi}(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$

Optimal convergence rate is achieved.

- Non-parametric case: Gaussian RKHS $\mathcal{H}$

$$\widehat{f} = \underset{g \in \mathcal{H}}{\arg\min} \left[\|g\|_{L^2}^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) + \frac{1}{n}\sum_{i'=1}^{n} g(\boldsymbol{x}'_{i'})\right) + \lambda \|g\|_{\mathcal{H}}^2\right]$$

$$n = n' \quad k_\gamma(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\gamma^2}\right)$$

For all $\rho, \rho' > 0$, there exists a constant K such that, for appropriately chosen λ, γ, the following inequality holds with probability $1 - 4e^{-\tau}$ for all $\tau, n \geq 1$:

$$\|\widehat{f} - f\|_{L^2}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}}^2 \leq K\left(n^{-\frac{2\alpha}{2\alpha+d}+\rho} + \tau n^{-1+\rho'}\right)$$

$\alpha \geq 0$: Regularity of Besov space that contains true f $\qquad \boldsymbol{x} \in \mathbb{R}^d$

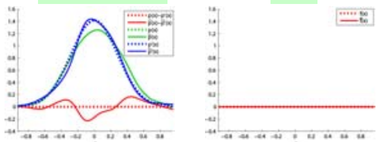
Optimal convergence rate is achieved.

Cf. Eberts & Steinwart (NIPS2011)

3.3 Numerical Examples

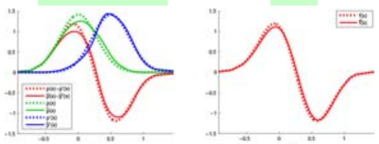
- $p(x) = p'(x) = N(x; 0, (4\pi)^{-1}) \qquad n = n' = 200$

Difference of KDEs — LSDD



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

Difference of KDEs — LSDD



4.1 L^2-Distance Approximation

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

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

- Naïve approximators via LSDD:
  - $L^2(p,p') = \int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} \approx \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H}\widehat{\boldsymbol{\theta}}$
  - $L^2(p,p') = \int f(\boldsymbol{x})\left(p(\boldsymbol{x}) - p'(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \approx \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}}$
- Consider their linear combination:

$$\beta \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} + (1-\beta)\widehat{\boldsymbol{\theta}}^\top \boldsymbol{H}\widehat{\boldsymbol{\theta}}$$

- For small λ, Taylor expansion gives

$$\beta \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} + (1-\beta)\widehat{\boldsymbol{\theta}}^\top \boldsymbol{H}\widehat{\boldsymbol{\theta}}$$
$$= \widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-1}\widehat{\boldsymbol{h}} - \lambda(2-\beta)\widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-2}\widehat{\boldsymbol{h}} + o_p(\lambda)$$

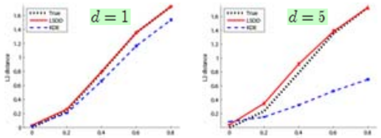
- Bias caused by regularization can be eliminated by $\beta = 2$.

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

4.2 Numerical Examples

- L^2-Distance Approximation $\qquad n = n' = 100$

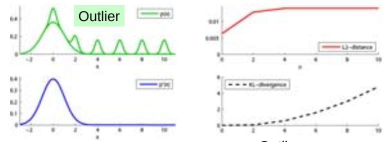
$$p(\boldsymbol{x}) = N(\boldsymbol{x}; (\mu, 0, \ldots, 0)^\top, (4\pi)^{-1}\boldsymbol{I}_d)$$
$$p'(\boldsymbol{x}) = N(\boldsymbol{x}; (0, 0, \ldots, 0)^\top, (4\pi)^{-1}\boldsymbol{I}_d)$$



- 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 more robust than KL-divergence.

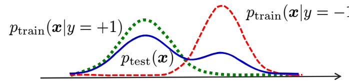
Cf. Basu, Harris, Hjort & Jones (Biometrika1998)

5.1 Class-Balance Estimation

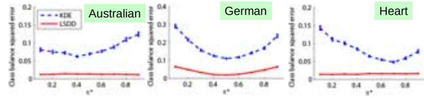
- Pattern recognition when class balances are different in training and test phases.
- If test class-balance is known, weighted learning eliminates estimation bias.
- When test class-balance is unknown, fit mixture of training class-wise input densities to test input density: Du Plessis & Sugiyama (ICML2012)

$$\min_{\pi \in [0,1]} L^2\left(p_{\mathrm{test}}(\boldsymbol{x}), q_\pi(\boldsymbol{x})\right)$$

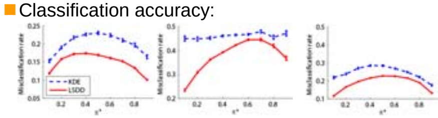
$$q_\pi(\boldsymbol{x}) = \pi p_{\mathrm{train}}(\boldsymbol{x}|y=+1) + (1-\pi) p_{\mathrm{train}}(\boldsymbol{x}|y=-1)$$



- Class-balance estimation error $\quad \pi^*$: True test class-balance



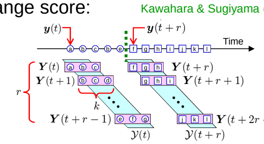
- Classification accuracy:



Regularized kernel least-squares classifier with class-balance weighting is used.

5.2 Change Detection

- Goal: Find change points in time-series.
- Use L^2-distance between past and current data as change score: Kawahara & Sugiyama (SADM2012)



CENSREC Speech Data — HASC Accelerometer Data