MLSS2012, Kyoto, Japan Sep. 7, 2012

Density Ratio Estimation in Machine Learning

Masashi Sugiyama
Tokyo Institute of Technology, Japan
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

Generative Approach to Machine Learning (ML)

- All ML tasks can be solved if data generating probability distributions are identified.

Knowing data generating distributions

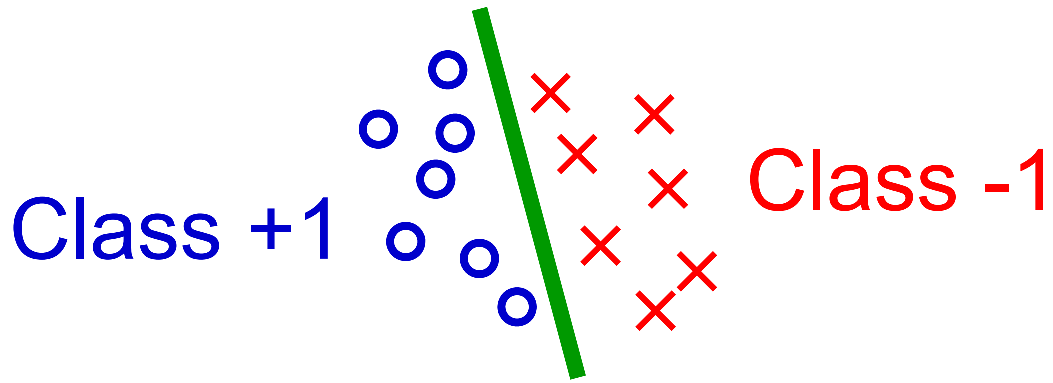
Knowing anything about data

- Thus, distribution estimation is the most general approach to ML.
- However, distribution estimation is hard without prior knowledge (i.e., non-parametric methods).

Discriminative Approach to ML

- Alternative approach: Solving a target ML task directly without distribution estimation.
- Ex: Support vector machine (SVM)
  - Without estimating data generating distributions, SVM directly learns a decision boundary.

Cortes & Vapnik (ML1995)



Discriminative Approach to ML

- However, there exist various ML tasks:
  - Learning under non-stationarity, domain adaptation, multi-task learning, two-sample test, outlier detection, change detection in time series, independence test, feature selection, dimension reduction, independent component analysis, causal inference, clustering, object matching, conditional probability estimation, probabilistic classification
- For each task, developing an ML algorithm that does not include distribution estimation is cumbersome/difficult.

Density-Ratio Approach to ML

- All ML tasks listed in the previous page include multiple probability distributions.

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

- For solving these tasks, individual densities are actually not necessary, but only the ratio of probability densities is enough:

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

- We directly estimate the density ratio without going through density estimation.

Intuitive Justification

Vapnik's principle: Vapnik (1998)

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

Knowing densities		Knowing ratio
$p(\boldsymbol{x}), q(\boldsymbol{x})$	$\Longrightarrow$ $\cancel{\Longleftarrow}$	$r(\boldsymbol{x}) = \dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$

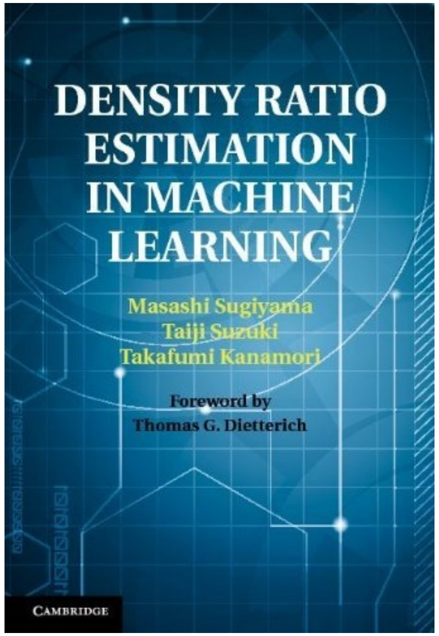
- Estimating the density ratio is substantially easier than estimating densities!

Quick Conclusions

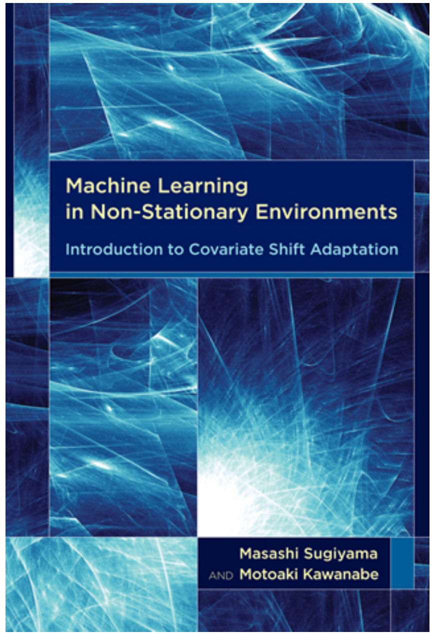
- Simple kernel least-squares (KLS) approach allows accurate and computationally efficient estimation of density ratios!
- Many ML tasks can be solved just by KLS:
  - Importance sampling: $\sum_{i=1}^{n} \frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \mathrm{loss}(\boldsymbol{x}_i)$
  - KL divergence estimation: $\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$
  - 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: $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})}$

Books on Density Ratios

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



- Sugiyama & Kawanabe
  Machine Learning
  in Non-Stationary Environments,
  MIT Press, 2012



Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
4. More on Density Ratio Estimation
5. Conclusions

Density Ratio Estimation: Problem Formulation

- Goal: Estimate the density ratio

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

from data

$$\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}} \overset{i.i.d.}{\sim} p_{\mathrm{nu}}(\boldsymbol{x})$$

$$\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}} \overset{i.i.d.}{\sim} p_{\mathrm{de}}(\boldsymbol{x})$$

Density Estimation Approach

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

$$\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}} \overset{i.i.d.}{\sim} p_{\mathrm{nu}}(\boldsymbol{x})$$

$$\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}} \overset{i.i.d.}{\sim} p_{\mathrm{de}}(\boldsymbol{x})$$

- Naïve 2-step approach:
  1. Perform density estimation:

     $$\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}), \widehat{p}_{\mathrm{de}}(\boldsymbol{x})$$

  2. Compute the ratio of estimated densities:

     $$\widehat{r}(\boldsymbol{x}) = \frac{\widehat{p}_{\mathrm{nu}}(\boldsymbol{x})}{\widehat{p}_{\mathrm{de}}(\boldsymbol{x})}$$

- However, this works poorly because 1. is performed without regard to 2.

Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
   - A) Probabilistic Classification
   - B) Moment Matching
   - C) Density Fitting
   - D) Density-Ratio Fitting
3. Usage of Density Ratios
4. More on Density Ratio Estimation
5. Conclusions

Probabilistic Classification

Qin (Biometrika1998), Bickel, Brückner & Scheffer (ICML2007)

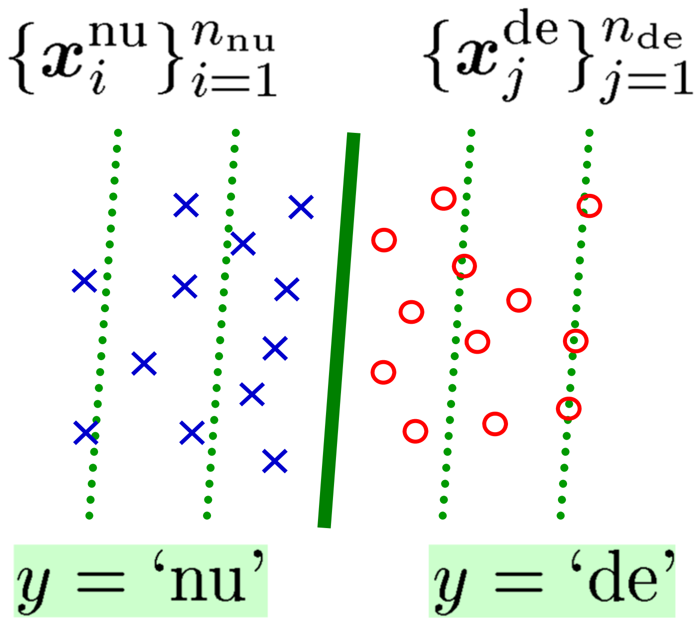
- Idea: Separate numerator and denominator samples by a probabilistic classifier.
- Via Bayes theorem

$$p(\boldsymbol{x}|y) = \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)}$$

density ratio is given by

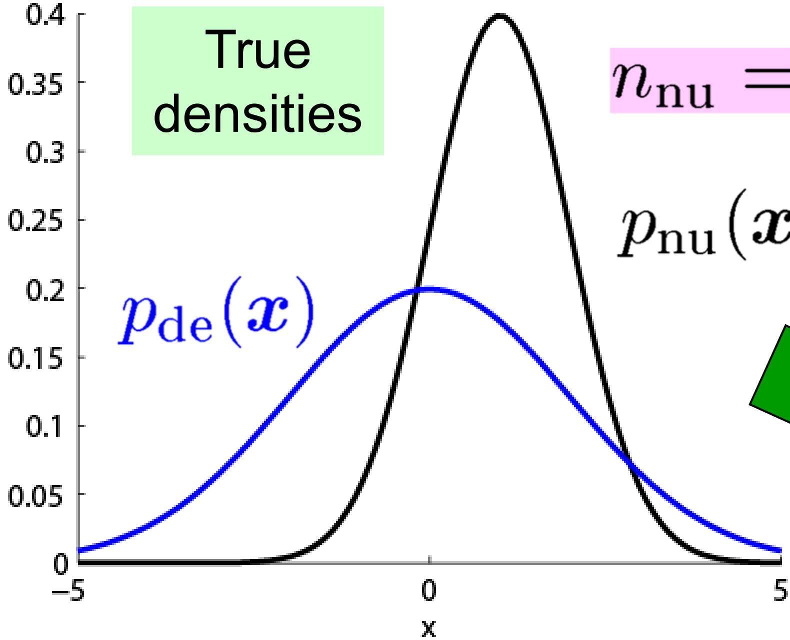
$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\mathrm{nu})}{p(\boldsymbol{x}|\mathrm{de})}$$

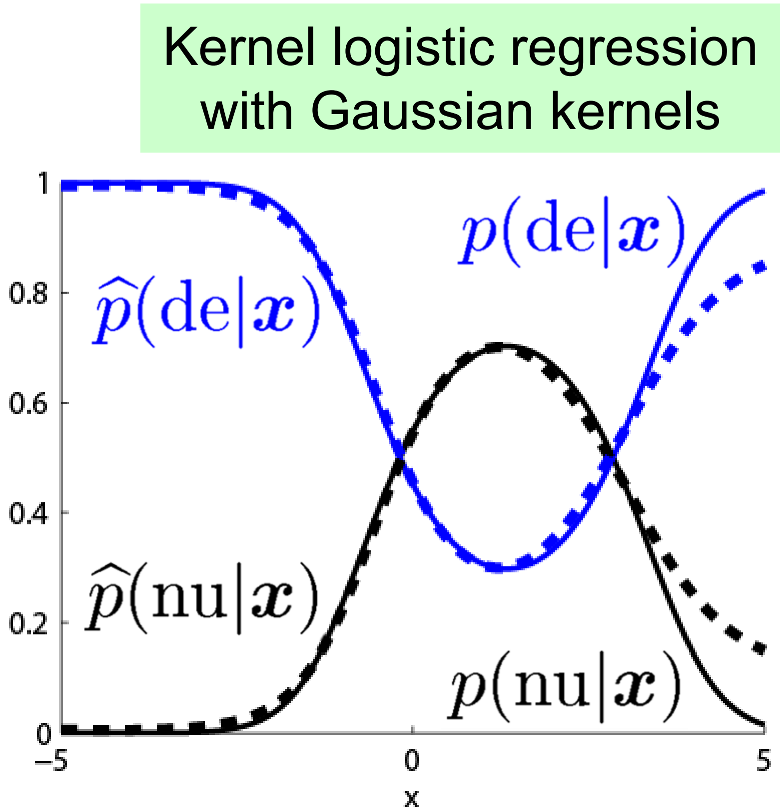
$$= \frac{p(\mathrm{de})}{p(\mathrm{nu})}\frac{p(\mathrm{nu}|\boldsymbol{x})}{p(\mathrm{de}|\boldsymbol{x})}$$

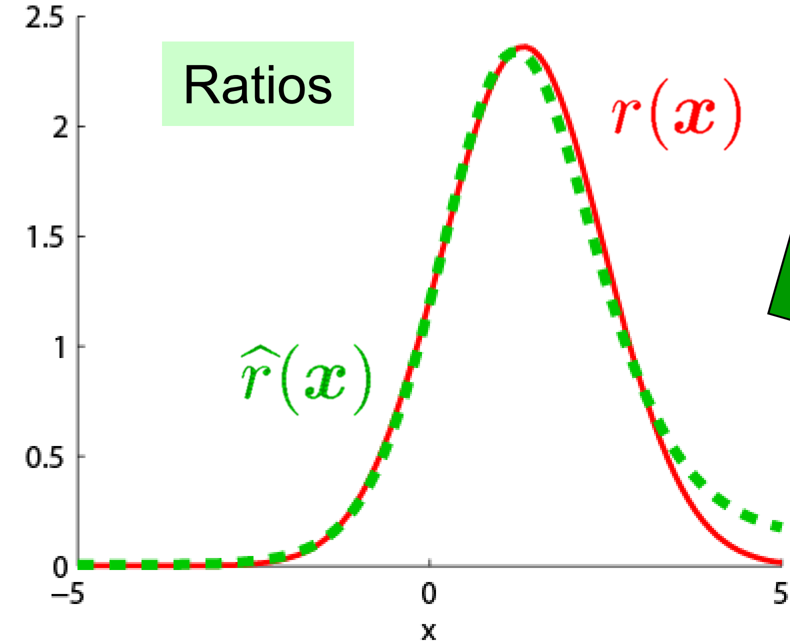


Numerical Example

$n_{\mathrm{nu}} = n_{\mathrm{de}} = 200$







$$\widehat{r}(\boldsymbol{x}) = \frac{\widehat{p}(\mathrm{de})}{\widehat{p}(\mathrm{nu})} \frac{\widehat{p}(\mathrm{nu}|\boldsymbol{x})}{\widehat{p}(\mathrm{de}|\boldsymbol{x})}$$

Probabilistic Classification: Summary

- Off-the-shelf software can be directly used.
- Logistic regression achieves the minimum asymptotic variance for correctly specified models.
  Qin (Biometrika1998)
  - However, not reliable for misspecified models.
    Kanamori, Suzuki & MS (IEICE2010)
- Multi-class classification gives density ratio estimates among multiple densities.
  Bickel, Bogojeska, Lengauer & Scheffer (ICML2008)

Organization of This Lecture

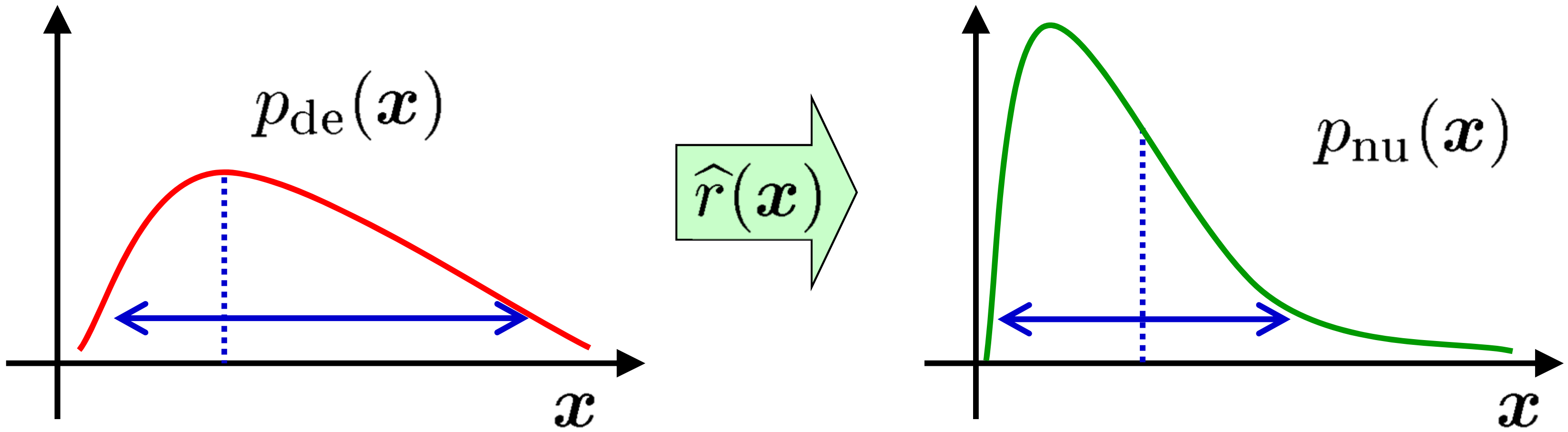
1. Introduction
2. Methods of Density Ratio Estimation
   - A) Probabilistic Classification
   - B) Moment Matching
   - C) Density Fitting
   - D) Density-Ratio Fitting
3. Usage of Density Ratios
4. More on Density Ratio Estimation
5. Conclusions

Moment Matching

Qin (Biometrika1998)

- Idea: Match moments of
  $\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x})p_{\mathrm{de}}(\boldsymbol{x})$ and $p_{\mathrm{nu}}(\boldsymbol{x})$.

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$



- Ex. Matching the mean:

$$\int \boldsymbol{x}\ \widehat{r}(\boldsymbol{x})p_{\mathrm{de}}(\boldsymbol{x})d\boldsymbol{x} = \int \boldsymbol{x}\ p_{\mathrm{nu}}(\boldsymbol{x})d\boldsymbol{x}$$

Moment Matching with Kernels

- Matching a finite number of moments does not necessarily yield the true density ratio even asymptotically.

- Kernel mean matching: All moments are efficiently matched in Gaussian RKHS $\mathcal{H}$:

Huang, Smola, Gretton, Borgwardt & Schölkopf (NIPS2006)

$$\min_{\widehat{r}\in\mathcal{H}} \left\| \int K(\boldsymbol{x},\cdot)\widehat{r}(\boldsymbol{x})p_{\mathrm{de}}(\boldsymbol{x})d\boldsymbol{x} - \int K(\boldsymbol{x},\cdot)p_{\mathrm{nu}}(\boldsymbol{x})d\boldsymbol{x} \right\|_{\mathcal{H}}^{2}$$

$K(\boldsymbol{x},\boldsymbol{x}')$:Gaussian kernel

Kernel Mean Matching

- Empirical optimization problem:

$$\min_{\beta_1,\ldots,\beta_{n_{\mathrm{de}}}} \quad \frac{1}{2}\sum_{j,j'=1}^{n_{\mathrm{de}}} \beta_j \beta_{j'} K(\boldsymbol{x}_j^{\mathrm{de}}, \boldsymbol{x}_{j'}^{\mathrm{de}}) - \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \sum_{j=1}^{n_{\mathrm{de}}} \beta_j \sum_{i=1}^{n_{\mathrm{nu}}} K(\boldsymbol{x}_i^{\mathrm{nu}}, \boldsymbol{x}_j^{\mathrm{de}})$$

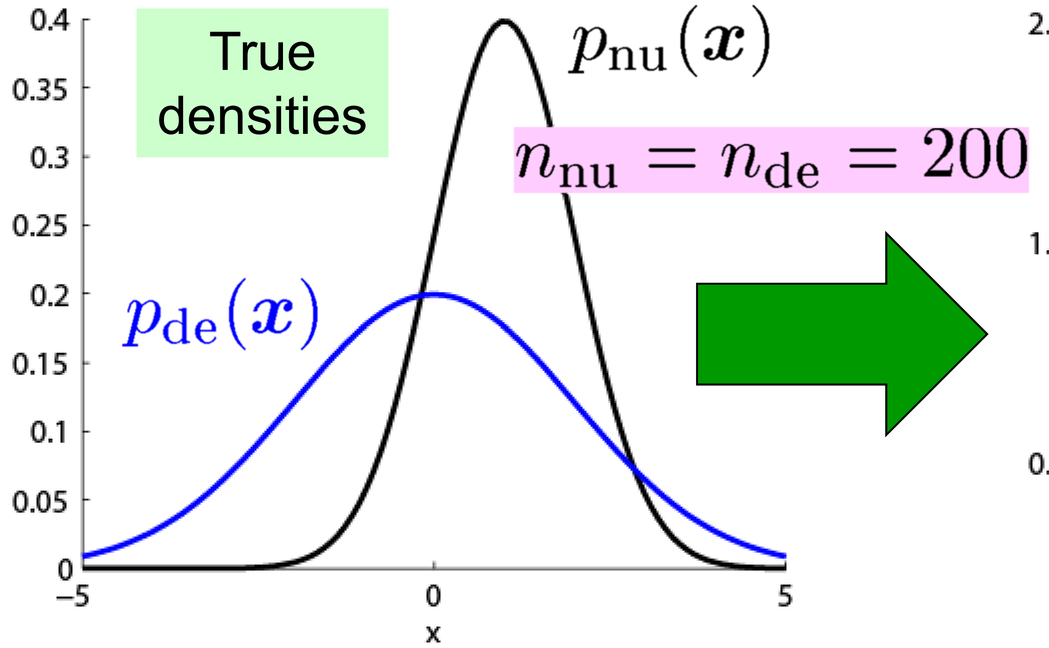
$$\text{subject to} \quad 0 \le \beta_1, \ldots, \beta_{n_{\mathrm{de}}} \le B \text{ and } \left| \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \beta_j - 1 \right| \le \epsilon$$

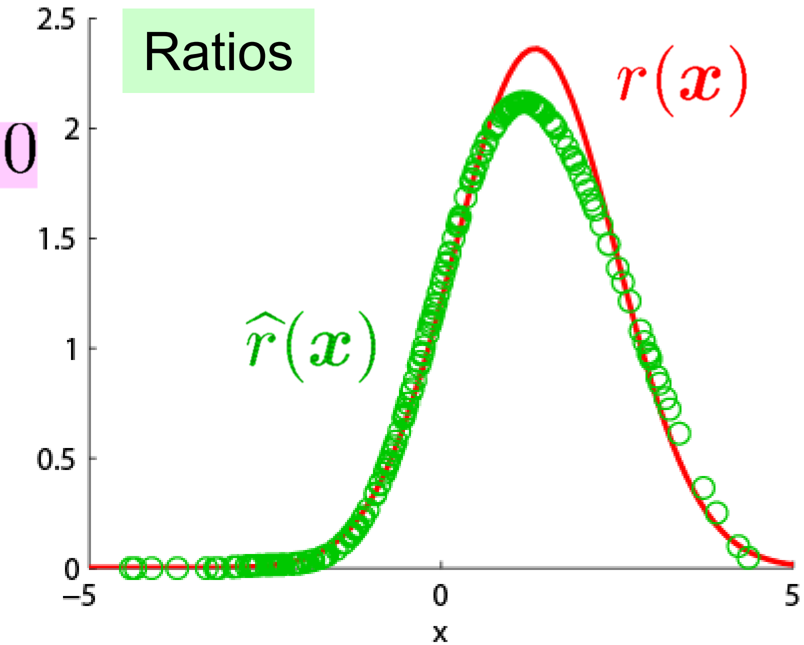
$K(\boldsymbol{x}, \boldsymbol{x}')$:Gaussian kernel

- This is a convex quadratic program.
- The solution directly gives density ratio estimates:

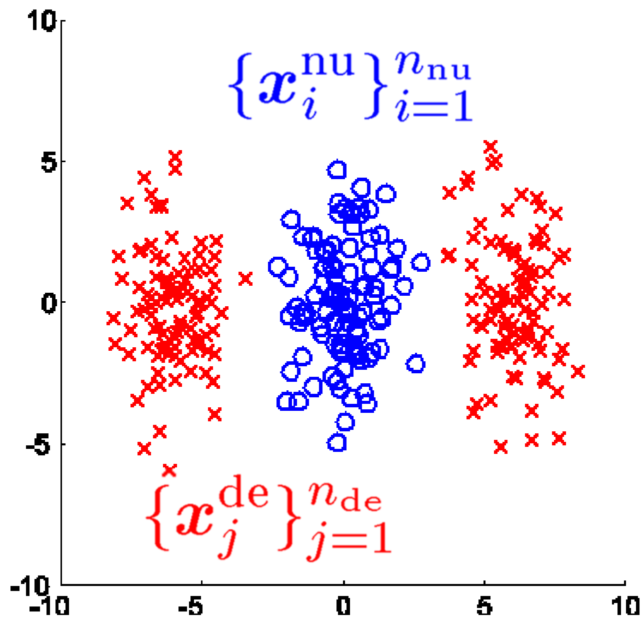
$$\widehat{\beta}_j = \widehat{r}(\boldsymbol{x}_j^{\mathrm{de}})$$

Numerical Example





- Kernel mean matching works well, given that the Gaussian width is appropriately chosen.
- A heuristic is to use the median distance between samples, but it may fail in a multi-modal case.



Moment Matching: Summary

- Finite moment matching is not consistent.
- Infinite moment matching with kernels:
  - Consistent and computationally efficient.
  - A convergence proof exists for reweighted means.

Gretton, Smola, Huang, Schmittfull, Borgwardt & Schölkopf (InBook 2009)

- Kernel parameter selection is cumbersome:
  - Changing kernels means changing error metrics.
  - Using the median distance between samples as the Gaussian width is a practical heuristic.
- A variant for learning the entire ratio function under general losses is also available.

Kanamori, Suzuki & MS (MLJ2012)

Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
   - A) Probabilistic Classification
   - B) Moment Matching
   - C) Density Fitting
   - D) Density-Ratio Fitting
3. Usage of Density Ratios
4. More on Density Ratio Estimation
5. Conclusions

Kullback-Leibler Importance Estimation Procedure (KLIEP)

Nguyen, Wainwright & Jordan (NIPS2007)
MS, Nakajima, Kashima, von Bünau & Kawanabe (NIPS2007)

- Minimize KL divergence from $p_{\mathrm{nu}}(\boldsymbol{x})$ to $\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x})$:

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

$$\min_{\widehat{r}} \underbrace{\int p_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{\widehat{r}(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x})} d\boldsymbol{x}}_{=:\mathrm{KL}(\widehat{r})}$$

- Decomposition of KL:

$$\mathrm{KL}(\widehat{r}) = C - \int p_{\mathrm{nu}}(\boldsymbol{x}) \log \widehat{r}(\boldsymbol{x}) d\boldsymbol{x}$$

Formulation

- Objective function:

$$\max_{\widehat{r}} \int p_{\mathrm{nu}}(\boldsymbol{x}) \log \widehat{r}(\boldsymbol{x}) d\boldsymbol{x}$$

- Constraints:
  - $\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x})$ is a probability density:

$$\int \widehat{r}(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x} = 1 \qquad \widehat{r}(\boldsymbol{x}) \geq 0$$

- Linear-in-parameter density-ratio model:

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}) \qquad \phi_\ell(\boldsymbol{x}) \geq 0$$

(ex. Gauss kernel)

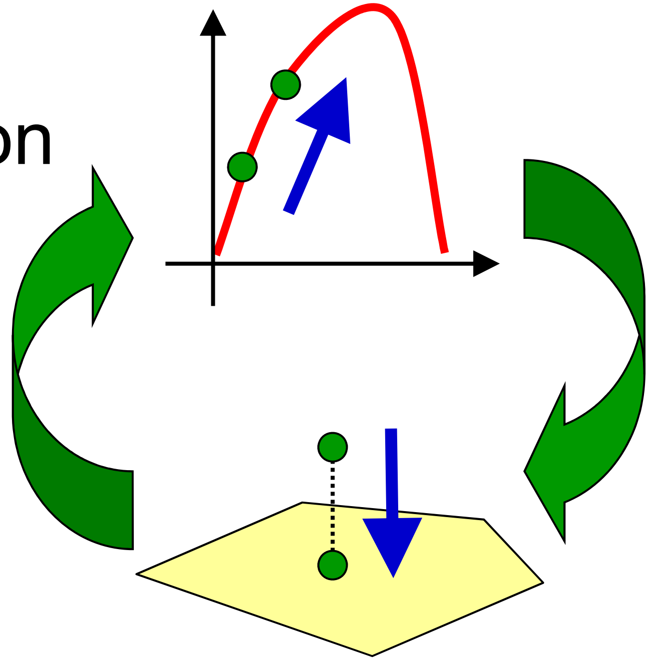
Algorithm

- Approximate expectations by sample averages:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n_{\mathrm{nu}}} \log(\boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}_i^{\mathrm{nu}})) \quad \text{subject to } \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}}) = 1 \text{ and } \boldsymbol{\alpha} \geq \mathbf{0}$$

- This is convex optimization, so repeating
  - Gradient ascent
  - Projection onto the feasible region

  leads to the global solution.



- The global solution is sparse!

Convergence Properties

Nguyen, Wainwright & Jordan (IEEE-IT2010)
MS, Suzuki, Nakajima, Kashima, von Bünau & Kawanabe (AISM2008)

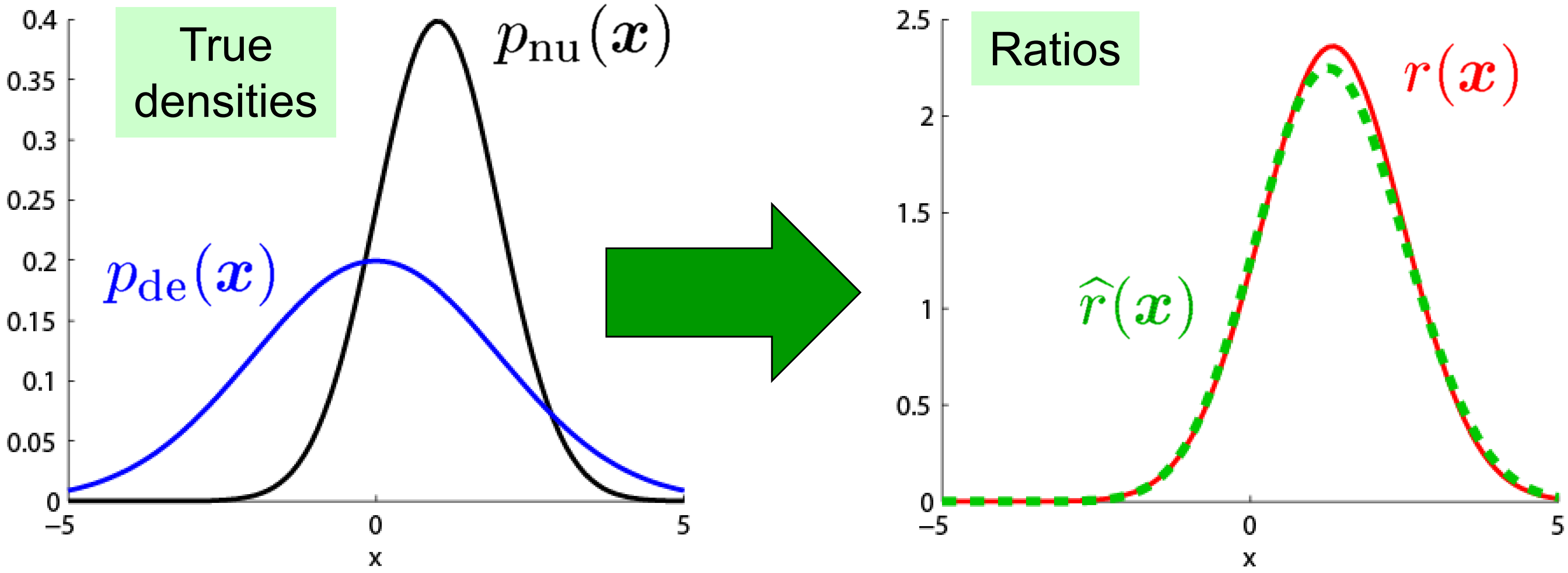
- **Parametric case:** $\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x})$
  - Learned parameter converge to the optimal value with order $n^{-\frac{1}{2}}$, which is the optimal rate.

    $n = \min(n_{\mathrm{nu}}, n_{\mathrm{de}})$
- **Non-parametric case:** $\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{n_{\mathrm{nu}}} \alpha_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell^{\mathrm{nu}})$
  - Learned function converges to the optimal function with order $n^{-\frac{1}{2+\gamma}}$, which is the optimal rate.

    $0 < \gamma < 2$: Complexity of the function class related to the covering number or bracketing entropy

Numerical Example

$$n_{\mathrm{nu}} = n_{\mathrm{de}} = 200$$



- Gaussian width can be determined by cross-validation with respect to KL.

$$\int p_{\mathrm{nu}}(\boldsymbol{x}) \log \widehat{r}(\boldsymbol{x}) d\boldsymbol{x}$$

Density Fitting under KL Divergence: Summary

- Cross-validation is available for kernel parameter selection.
- Variations for various models exist:
  - Log-linear, Gaussian mixture, PCA mixture, etc.
- Elaborate ratios such as $\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$ can also be estimated.
- An unconstrained variant corresponds to maximizing a lower-bound of KL divergence.

$$\int p_{\mathrm{nu}}(\boldsymbol{x}) \log \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$

Nguyen, Wainwright & Jordan (NIPS2007)

Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
   A) Probabilistic Classification
   B) Moment Matching
   C) Density Fitting
   D) Density-Ratio Fitting
3. Usage of Density Ratios
4. More on Density Ratio Estimation
5. Conclusions

Least-Squares Importance Fitting (LSIF)

Kanamori, Hido & MS (NIPS2008)

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

- Minimize squared-loss:

$$\min_{\widehat{r}} \underbrace{\int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x}}_{=:\mathrm{SQ}(\widehat{r})}$$

- Decomposition and approximation of SQ:

$$\mathrm{SQ}(\widehat{r}) = \int \left(\widehat{r}(\boldsymbol{x})\right)^2 p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x} - 2\int \widehat{r}(\boldsymbol{x}) p_{\mathrm{nu}}(\boldsymbol{x}) d\boldsymbol{x} + C$$

$$\approx \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \widehat{r}(\boldsymbol{x}_j^{\mathrm{de}})^2 - \frac{2}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}}) + C$$

Constrained Formulation

- Linear (or kernel) density-ratio model:

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

- Constrained LSIF (cLSIF):
  - Non-negativity constraint with ℓ_1-regularizer

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \mathbf{1} \right]$$

$$\text{subject to } \boldsymbol{\alpha} \geq \mathbf{0}$$

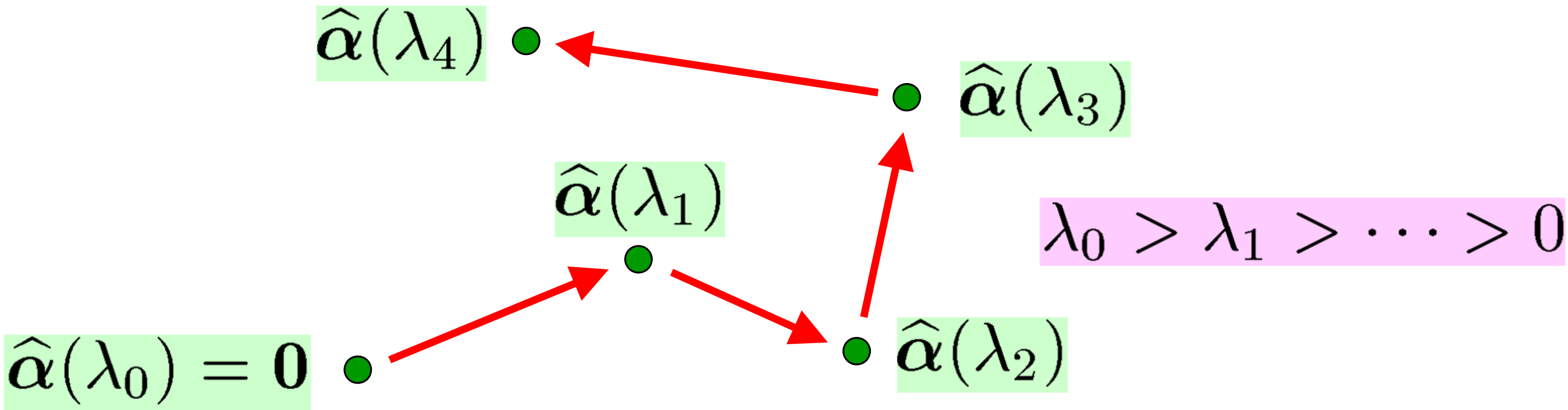
  - A convex quadratic program with sparse solution.

$$\widehat{\boldsymbol{H}} = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}}) \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}})^\top \qquad \widehat{\boldsymbol{h}} = \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{\phi}(\boldsymbol{x}_i^{\mathrm{nu}})$$

Regularization Path Tracking

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \mathbf{1} \right] \quad \text{subject to } \boldsymbol{\alpha} \geq \mathbf{0}$$

- The solution path is piece-wise linear with respect to the regularization parameter λ .



- Solutions for all λ can be computed efficiently without QP solvers!

Unconstrained Formulation

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

Unconstrained LSIF (uLSIF):

- uLSIF: No constraint with ℓ_2-regularizer

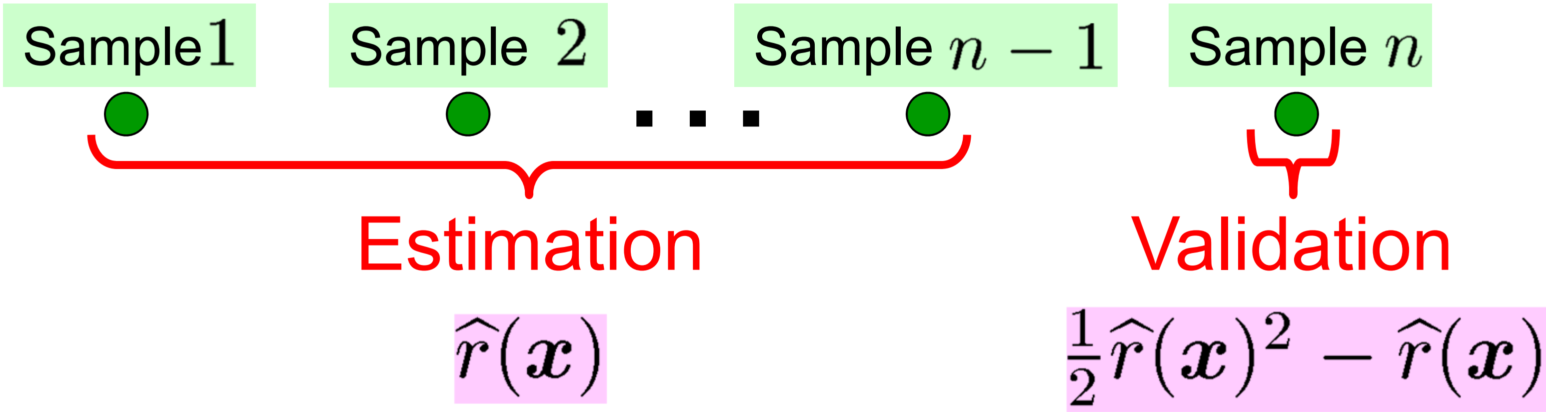
$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right]$$

- Analytic solution is available: $(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$

$$\widehat{\boldsymbol{H}} = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}}) \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}})^\top \qquad \widehat{\boldsymbol{h}} = \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{\phi}(\boldsymbol{x}_i^{\mathrm{nu}})$$

Analytic LOOCV Score

- Leave-out-one cross-validation (LOOCV):



- LOOCV generally requires n repetitions.
- However, it can be analytically computed for uLSIF (Sherman-Woodbury-Morrison formula).
- Computation time including model selection is significantly reduced.

Theoretical Properties of uLSIF

Parametric convergence:

- Learned parameter converge to the optimal value with order $n^{-\frac{1}{2}}$, which is the optimal rate.

$n = \min(n_{\mathrm{nu}}, n_{\mathrm{de}})$ Kanamori, Hido & MS (JMLR2009)

Non-parametric convergence:

- Learned function converges to the optimal function with order $n^{-\frac{1}{2+\gamma}}$ (depending on the bracketing entropy), which is the optimal rate.

$0 < \gamma < 2$ Kanamori, Suzuki & MS (MLJ2012)

Non-parametric numerical stability:

- uLSIF has the smallest condition number among a class of density ratio estimators.

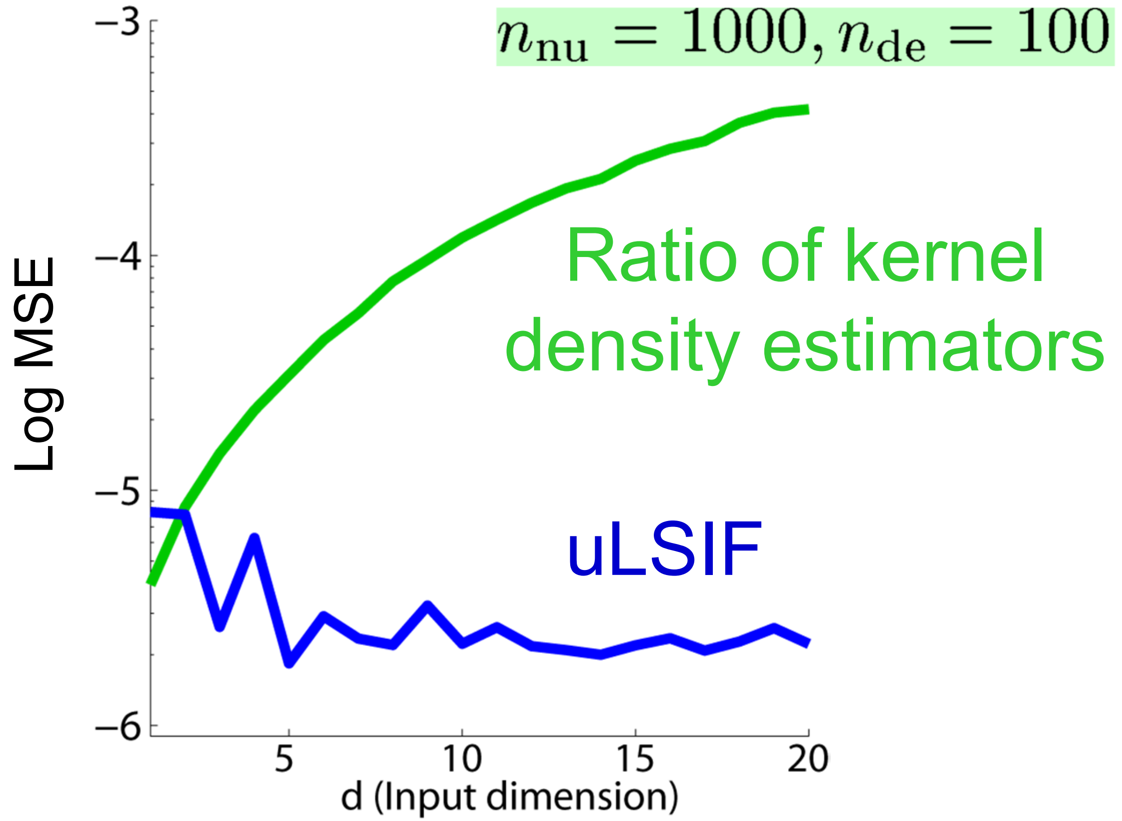
Kanamori, Suzuki & MS (ArXiv2009)

Numerical Example

$$p_{\mathrm{nu}}(\boldsymbol{x}) = N(\boldsymbol{x}; (0,0,\ldots,0)^\top, \boldsymbol{I}_d)$$

$$p_{\mathrm{de}}(\boldsymbol{x}) = N(\boldsymbol{x}; (1,0,\ldots,0)^\top, \boldsymbol{I}_d)$$

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$



Density-Ratio Fitting: Summary

- LS formulation is computationally efficient:
  - cLSIF: Regularization path tracking
  - uLSIF: Analytic solution and LOOCV
- Gives an accurate approximator of Pearson (PE) divergence (an *f*-divergence):

$$\int p_{\mathrm{de}}(\boldsymbol{x}) \left(\frac{p_{\mathrm{nu}}(\boldsymbol{x})}{q_{\mathrm{de}}(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$$

- Analytic solution of uLSIF allows us to compute the derivative of PE divergence approximator:
  - Useful in dimension reduction, independent component analysis, causal inference etc.

Qualitative Comparison of Density Ratio Estimation Methods

	Density estimation	Computation cost	Elaborate ratio estimation	Cross validation	Model flexibility
Probabilistic classification	Avoided	$n_{\mathrm{nu}} + n_{\mathrm{de}}$ parameters learned by quasi Newton	Not possible	Possible	Kernel
Moment matching	Avoided	n_{nu} parameters learned by QP	Not possible	Not possible	Kernel
Density fitting	Avoided	n_{nu} parameters learned by gradient and projection	Possible	Possible	Kernel, log-kernel, Gauss-mix, PCA-mix
Density ratio fitting	Avoided	n_{nu} parameters learned analytically	Possible	Possible	Kernel

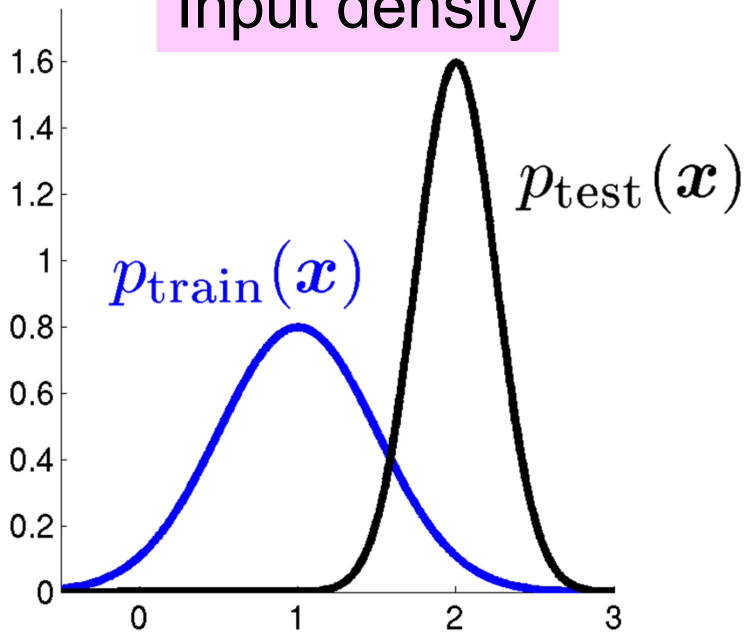
Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
   A) Importance sampling
   B) Distribution comparison
   C) Mutual information estimation
   D) Conditional probability estimation
4. More on Density Ratio Estimation
5. Conclusions

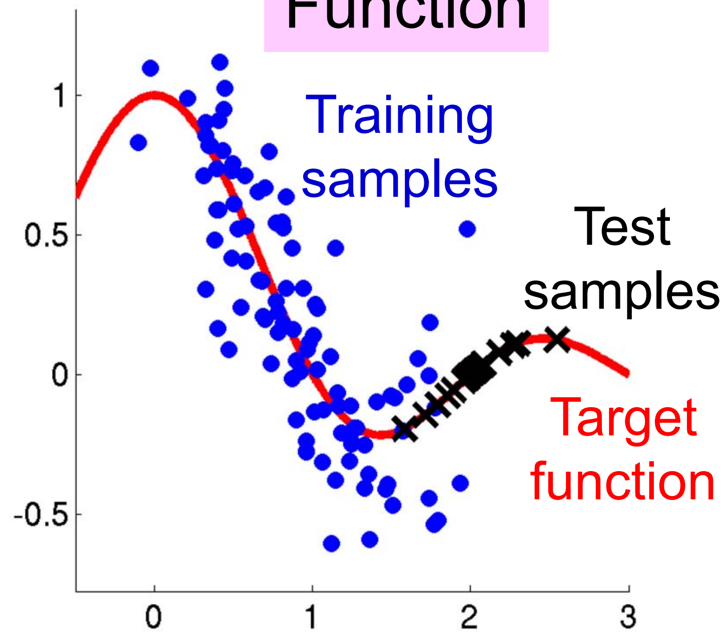
Learning under Covariate Shift

- **Covariate shift:** Shimodaira (JSPI2000)
  - Training/test input distributions are different, but target function remains unchanged.
  - (Weak) extrapolation.

Input density



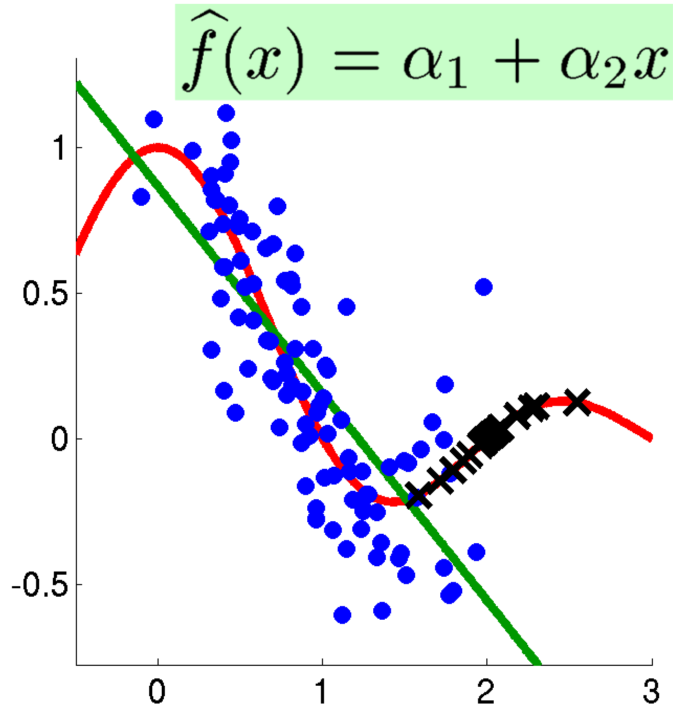
Function



Ordinary Least-Squares (OLS)

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

- In standard setting, OLS is consistent, i.e., the learned function converges to the best solution when $n \to \infty$.
- Under covariate shift, OLS is no longer consistent.



Law of Large Numbers

- Sample average converges to the population mean:

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{loss}(\boldsymbol{x}_i) \longrightarrow \int \mathrm{loss}(\boldsymbol{x})p_{train}(\boldsymbol{x})d\boldsymbol{x}$$

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} p_{train}(\boldsymbol{x})$$

- We want to estimate the expectation over test input points only using training input points $\{\boldsymbol{x}_i\}_{i=1}^{n}$.

$$\int \mathrm{loss}(\boldsymbol{x})p_{test}(\boldsymbol{x})d\boldsymbol{x}$$

Importance Weighting

- **Importance**: Ratio of test and training input densities

$$\frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}$$

- **Importance-weighted average**:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)}\mathrm{loss}(\boldsymbol{x}_i) \qquad \boldsymbol{x}_i \overset{i.i.d.}{\sim} p_{train}(\boldsymbol{x})$$

$$\longrightarrow \int \frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}\mathrm{loss}(\boldsymbol{x})p_{train}(\boldsymbol{x})d\boldsymbol{x}$$

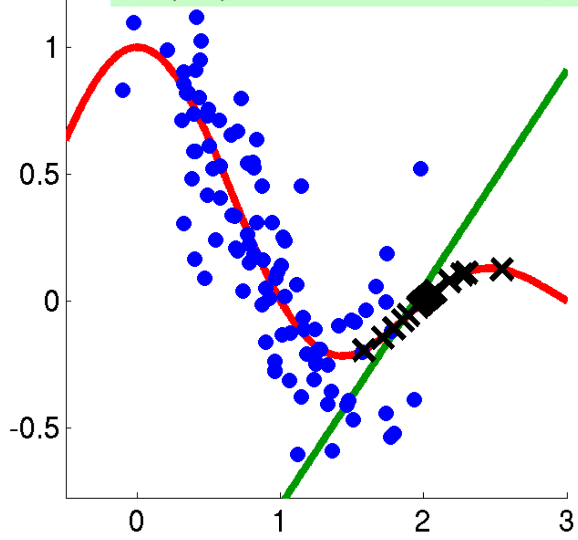
$$= \int \mathrm{loss}(\boldsymbol{x})p_{test}(\boldsymbol{x})d\boldsymbol{x}$$

Importance-Weighted Least-Squares

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

- IWLS is consistent even under covariate shift.
- The idea is applicable to any likelihood-based methods!
  - Support vector machine, logistic regression, conditional random field, etc.

$$\widehat{f}(x) = \alpha_1 + \alpha_2 x$$



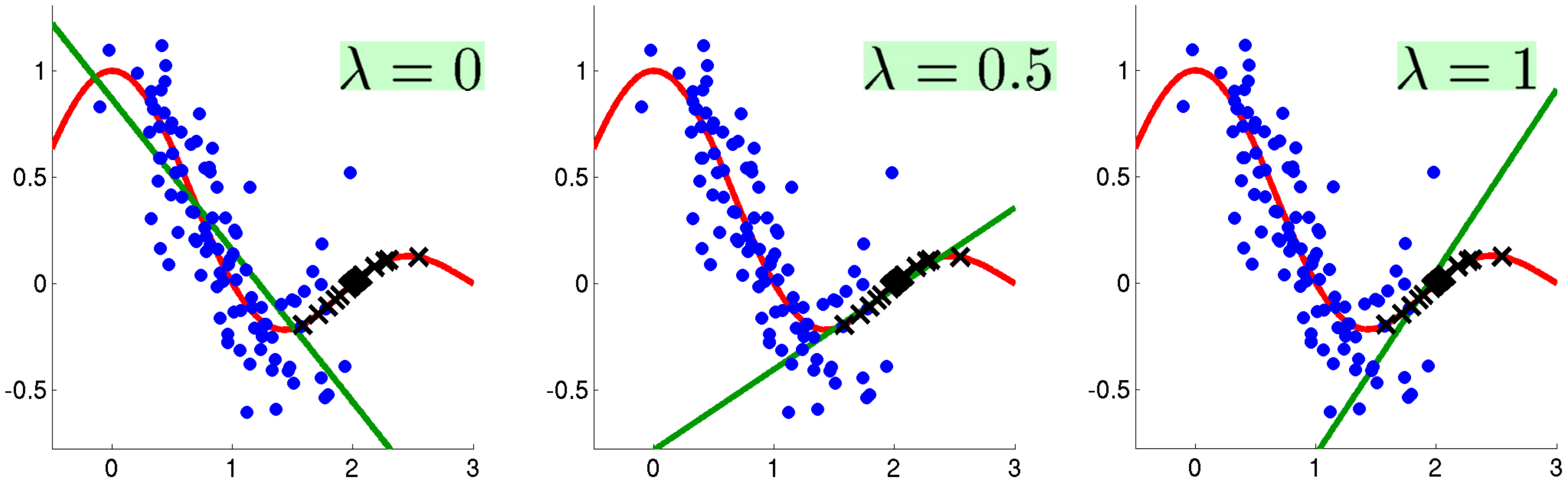
Model Selection

- Controlling bias-variance trade-off is important.
  - No weighting: low-variance, high-bias
  - Importance weighting: low-bias, high-variance

- "Flattened"-IWLS:

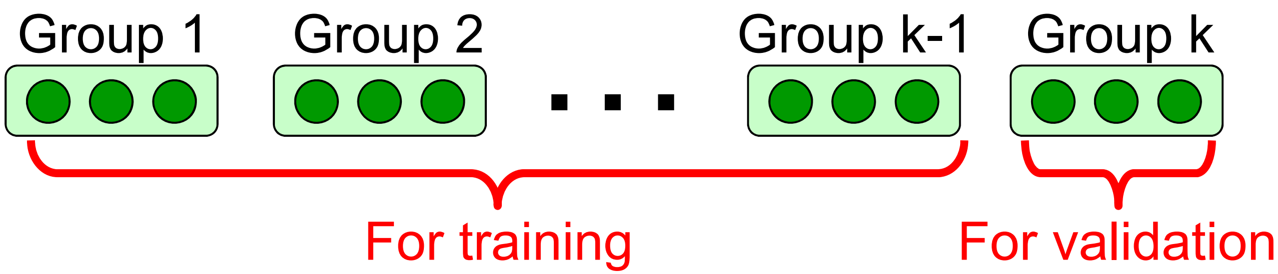
$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \right)^{\lambda} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

Shimodaira (JSPI2000)



Model Selection

- Importance weighting also plays a central role for unbiased model selection:
  - Akaike information criterion (regular models)
    Shimodaira (JSPI2000)
  - Subspace information criterion (linear models)
    MS & Müller (Stat&Dec.2005)
  - Cross-validation (arbitrary models)
    MS, Krauledat & Müller (JMLR2007)



$$\widehat{f}(\boldsymbol{x})$$

$$\frac{p_{test}(\boldsymbol{x}')}{p_{train}(\boldsymbol{x}')}\left(\widehat{f}(\boldsymbol{x}') - y'\right)^2$$

Experiments: Speaker Identification

Yamada, MS & Matsui (SigPro2010)

- NTT Japanese speech dataset: Matsui & Furui (ICASSP1993)
- Text-independent speaker identification accuracy for 10 male speakers.
- Kernel logistic regression (KLR) with sequence kernel.

Training data	Speech length	IWKLR+IWCV+KLIEP	KLR+CV
9 months before	1.5 [sec]	91.0 %	88.2 %
	3.0 [sec]	95.0 %	92.9 %
	4.5 [sec]	97.7 %	96.1 %
6 months before	1.5 [sec]	91.0 %	87.7 %
	3.0 [sec]	95.3 %	91.1 %
	4.5 [sec]	97.4 %	93.4 %
3 months before	1.5 [sec]	94.8 %	91.7 %
	3.0 [sec]	97.9 %	96.3 %
	4.5 [sec]	98.8 %	98.3 %

Experiments: Text Segmentation

Tsuboi, Kashima, Hido, Bickel & MS (JIP2009)

こんな失敗はご愛敬だよ.
→ こんな／失敗／は／ご／愛敬／だ／よ／.

- Japanese word segmentation dataset.

  Tsuboi, Kashima, Mori, Oda & Matsumoto (COLING2008)

- Adaptation from daily conversation to medical domain.
- Segmentation by conditional random field (CRF).

	IWCRF+IWCV +KLIEP	CRF+CV	CRF+CV (use additional test labels)
F-measure (larger is better)	94.46	92.30	94.43

Semi-supervised adaptation with importance weighting is comparable to supervised adaptation!

Other Applications

- Age prediction from faces:
  - Illumination change

    Ueki, MS & Ihara (ICPR2010)

- Brain-computer interface:
  - Mental condition change

    MS, Krauledat & Müller (JMLR2007)
    Li, Kambara, Koike & MS (IEEE-TBME2010)

- Robot control:
  - Efficient sample reuse

    Hachiya, Akiyama, MS & Peters (NN2009)
    Hachiya, Peters & MS (NeCo2011)

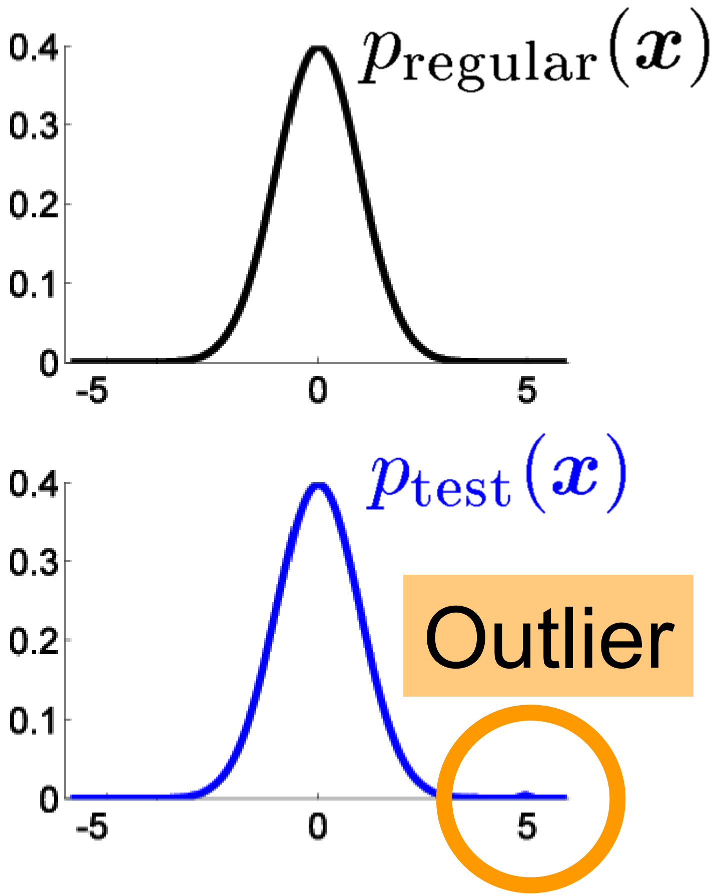
Organization of This Lecture

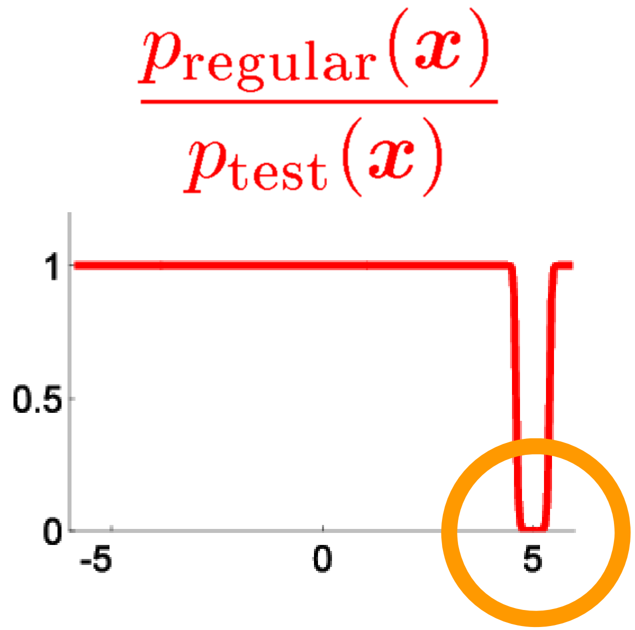
1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
   - A) Importance sampling
   - B) Distribution comparison
   - C) Mutual information estimation
   - D) Conditional probability estimation
4. More on Density Ratio Estimation
5. Conclusions

Inlier-Based Outlier Detection

Hido, Tsuboi, Kashima, MS & Kanamori (ICDM2008, KAIS2011)
Smola, Song & Teo (AISTATS2009)

- Goal: Given a set of inlier samples, find outliers in a test set (if exist)



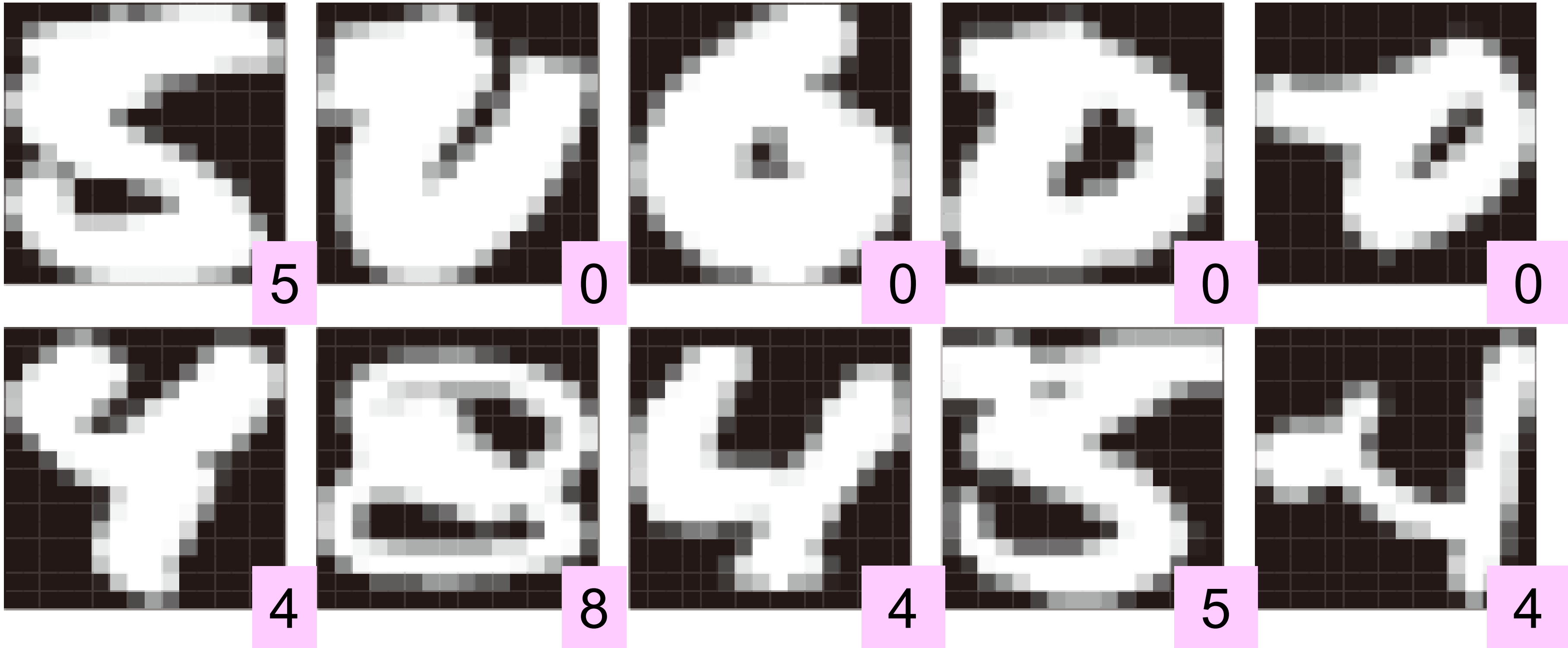


Tuning parameters can be optimized in terms of ratio approximation error

Experiments

Hido, Tsuboi, Kashima, MS & Kanamori (ICDM2008, KAIS2011)

- Top10 outliers in the USPS test dataset found based on the USPS training dataset.



Most of them are not readable even by human.

Failure Prediction in Hard-Disk Drives

- Self-Monitoring And Reporting Technology (SMART):

Murray, Hughes & Kreutz-Delgado (JMLR 2005)

	Least-squares density ratio	One-class SVM	Local outlier factor	
			#NN=5	#NN=30
AUC (larger is better)	0.881	0.843	0.847	0.924
Comp. time	1	26.98	65.31	

- LOF works well, given #NN is set appropriately. But there is no objective model selection method.
- Density ratio method can use cross-validation for model selection, and is computationally efficient.

OSVM: Schölkopf, Platt, Shawe-Taylor, Smola & Williamson (NeCo2001)
LOF: Breunig, Kriegel, Ng & Sander (SIGMOD2000)

Other Applications

- Steel plant diagnosis — Hirata, Kawahara & MS (Patent2011)
- Printer roller quality control — Takimoto, Matsugu & MS (DMSS2009)
- Loan customer inspection — Hido, Tsuboi, Kashima, MS & Kanamori (KAIS2011)
- Sleep therapy — Kawahara & MS (SADM2012)

Divergence Estimation

Nguyen, Wainwright & Jordan (IEEE-IT2010)
MS, Suzuki, Ito, Kanamori & Kimura (NN2011)

- **Goal**: Estimate a divergence functional from

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

  - Kullback-Leibler divergence: $\displaystyle\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$
  - Pearson divergence: $\displaystyle\int p'(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$
    (an *f*-divergence)

- Use density ratio estimation: $\displaystyle r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$

Real-World Applications

- Regions-of-interest detection in images:

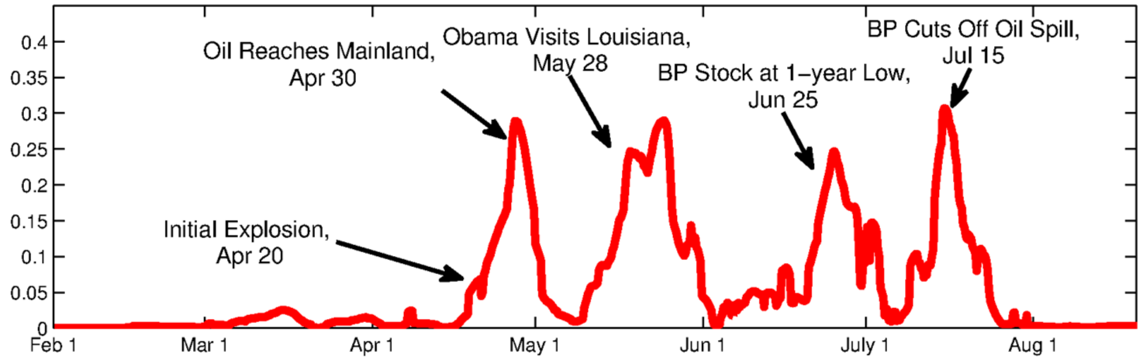
Yamanaka, Matsugu & MS
(IEEJ2011)

- Event detection in movies:

Matsugu, Yamanaka & MS
(VECTaR2011)

- Event detection from Twitter data:

Liu, Yamada, Collier
& MS (arXiv2012)



Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
   - A) Importance sampling
   - B) Distribution comparison
   - C) Mutual information estimation
   - D) Conditional probability estimation
4. More on Density Ratio Estimation
5. Conclusions

Mutual Information Estimation

Suzuki, MS, Sese & Kanamori (FSDM2008)

- Mutual information (MI): Shannon (1948)

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

- MI works as an independence measure:

$\mathrm{MI} = 0$ $\Longleftrightarrow$ $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent

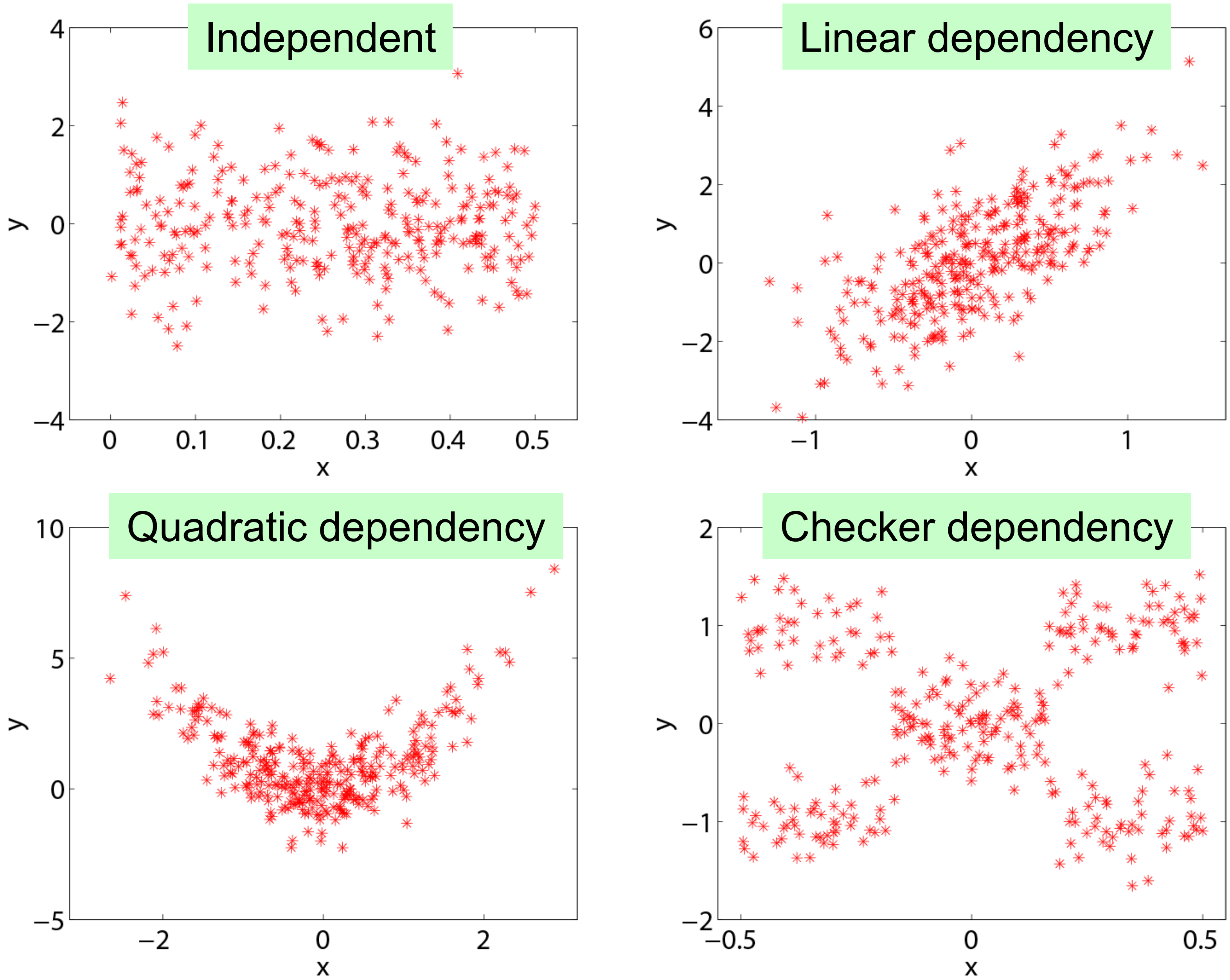
- Use KL-based density ratio estimation (KLIEP):

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

Experiments: Methods Compared

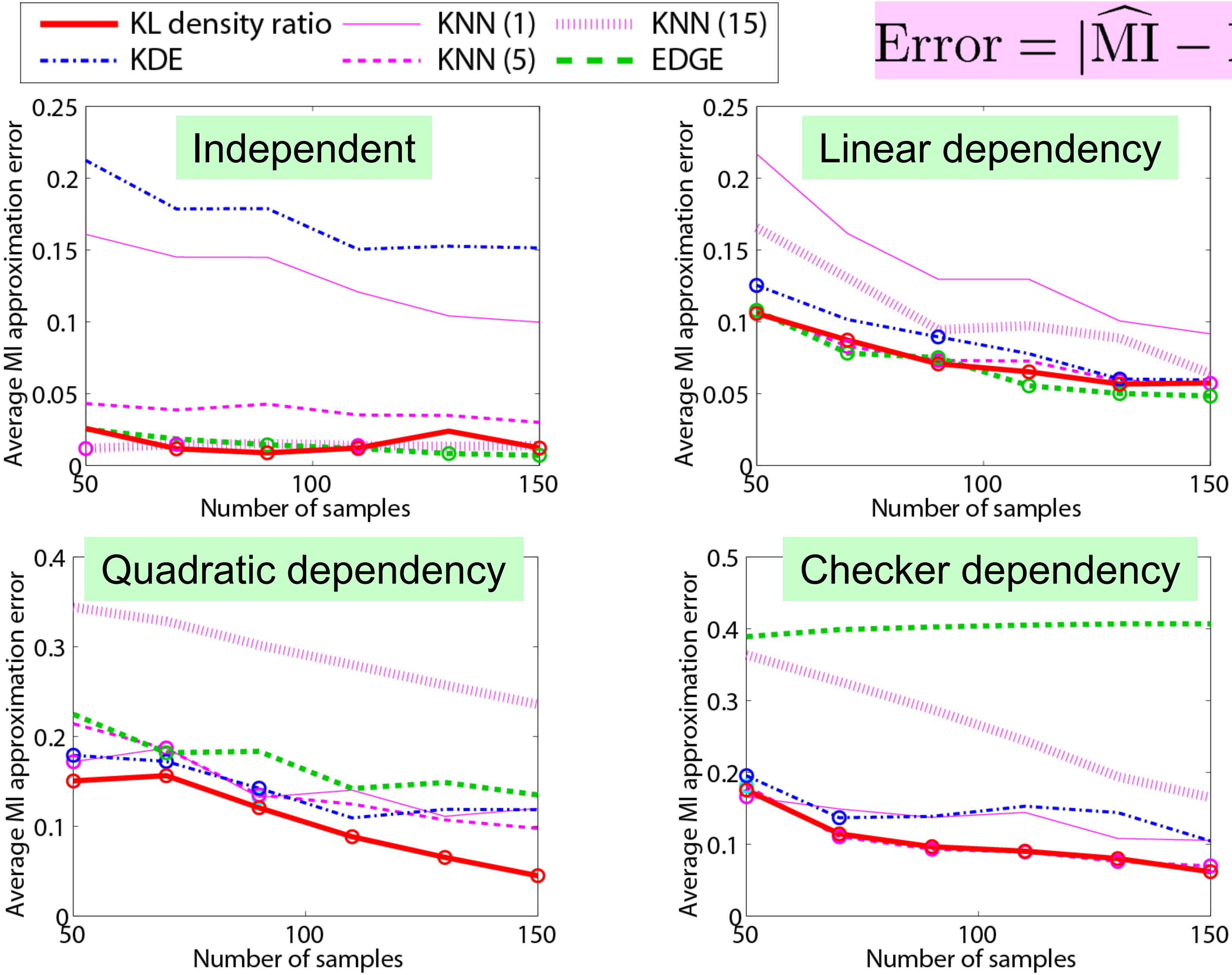
- KL-based density ratio method.
- Kernel density estimation (KDE).
- K-nearest neighbor density estimation (KNN). Kraskov, Stögbauer & Grassberger (PRE2004)
  - The number of NNs is a tuning parameter.
- Edgeworth expansion density estimation (EDGE). van Hulle (NeCo2005)

Datasets for Evaluation



MI Approximation Error

$$\text{Error} = |\widehat{\text{MI}} - \text{MI}|$$



Estimation of Squared-Loss Mutual Information (SMI)

Suzuki, MS, Sese & Kanamori (BMC Bioinfo. 2009)

- Ordinary MI is based on the KL-divergence.
- SMI is the Pearson divergence:

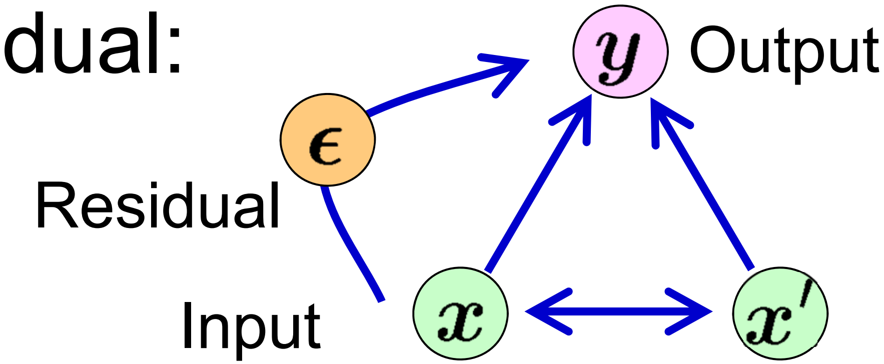
$$\mathrm{SMI} = \iint p(\boldsymbol{x})p(\boldsymbol{y}) \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

  - Can also be used as an independence measure.
  - Can be approximated analytically and efficiently by least-squares density ratio estimation (uLSIF).

Usage of SMI Estimator

- Between input and output:
  - Feature ranking — Suzuki, MS, Sese & Kanamori (BMCBioinfo 2009)
  - Sufficient dimension reduction — Suzuki & MS (NeCo2012)
  - Clustering — MS, Yamada, Kimura & Hachiya (ICML2011) Kimura & MS (JACIII2011)
- Between inputs:
  - Independent component analysis — Suzuki & MS (NeCo2010)
  - Object matching — Yamada & MS (AISTATS2011)
  - Canonical dependency analysis — Karasuyama & MS (NN2012)
- Between input and residual:
  - Causal inference

    Yamada & MS (AAAI2010)



Sufficient Dimension Reduction

Li (JASA1991)

- Input: $\boldsymbol{x}$
- Output: $\boldsymbol{y}$
- Projected input: $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$

$$\boldsymbol{z} = \boldsymbol{W} \boldsymbol{x}$$

$$\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}$$

- Goal: Find $\boldsymbol{W}$ so that $\boldsymbol{z}$ contains all information on $\boldsymbol{y}$, i.e., $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} \mid \boldsymbol{z}$
  - In terms of SMI: Suzuki & MS (NeCo2012)

$$\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} \mid \boldsymbol{z} \quad \Longleftrightarrow \quad \max_{\boldsymbol{W}} \ \mathrm{SMI}(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{y})$$

Sufficient Dimension Reduction via SMI Maximization

- Let's solve $\max_{\boldsymbol{W}} \ \widehat{\mathrm{SMI}}(\boldsymbol{W}) \text{ subject to } \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}$.

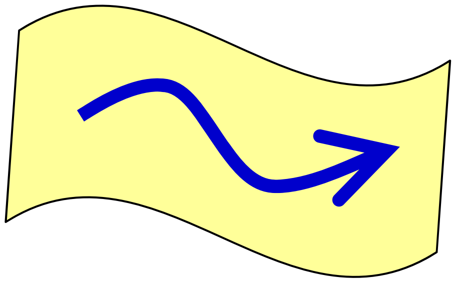
$$\widehat{\mathrm{SMI}}(\boldsymbol{W}) = 2\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{H}} \widehat{\boldsymbol{\alpha}} - 1$$

$\widehat{\boldsymbol{\alpha}}$: uLSIF solution

- Since $\boldsymbol{W}$ is on a Grassmann manifold, natural gradient gives the steepest direction:

Amari (NeCo1998)

$$\boldsymbol{W} \longleftarrow \boldsymbol{W} + \epsilon \frac{\partial \widehat{\mathrm{SMI}}}{\partial \boldsymbol{W}} \left(\boldsymbol{I} - \boldsymbol{W}^\top \boldsymbol{W} \right)$$



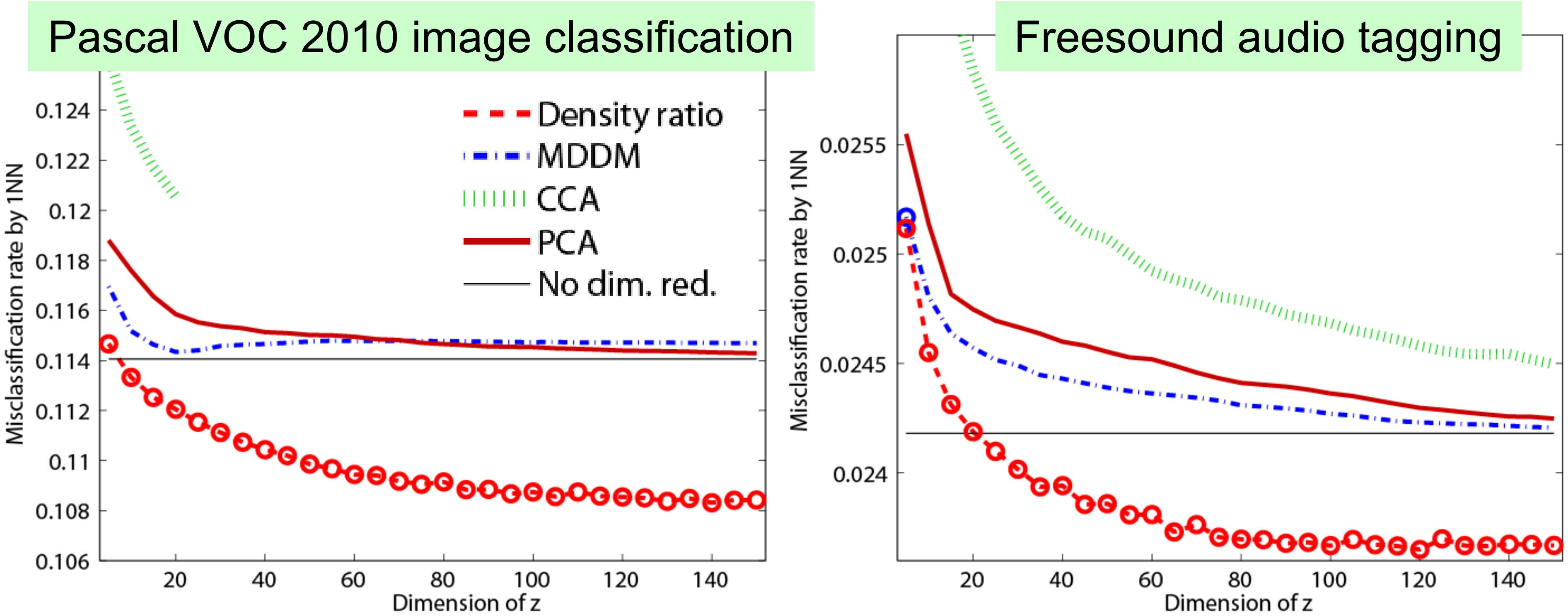
- A computationally efficient heuristic update is also available.

Yamada, Niu, Takagi & MS (ACML2011)

Experiments

Yamada, Niu, Takagi & MS (ACML2011)

Dimension reduction for multi-label data:



- MDDM: Multi-label dimensionality reduction via dependence maximization (MDDM) Zhang & Zhou (ACM-TKDD2010)
- CCA: Canonical correlation analysis
- PCA: Principal component analysis

Organization of This Lecture

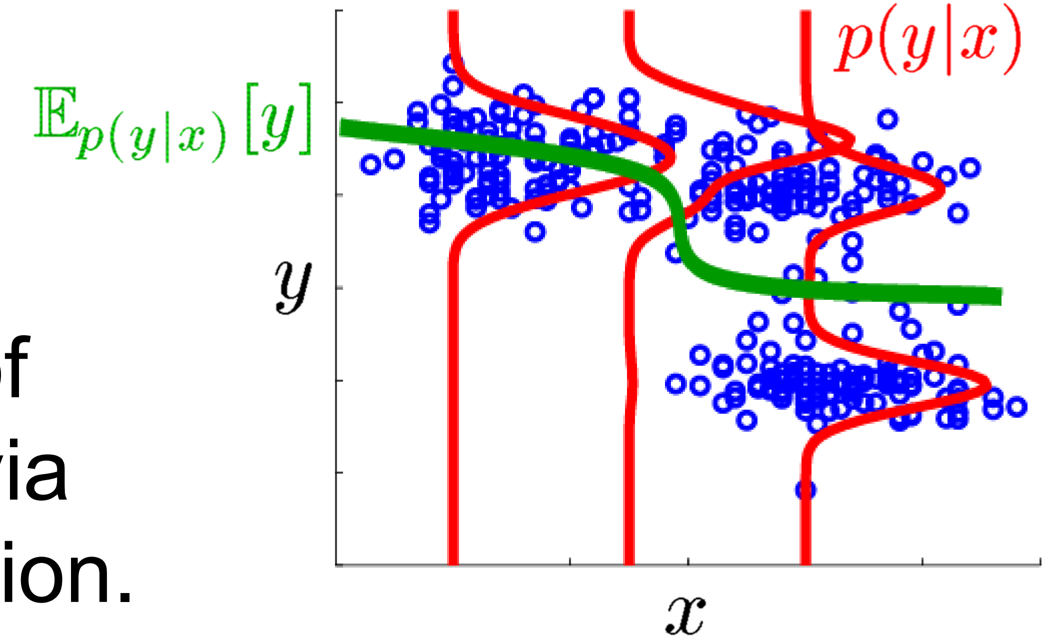
1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
   - A) Importance sampling
   - B) Distribution comparison
   - C) Mutual information estimation
   - D) Conditional probability estimation
4. More on Density Ratio Estimation
5. Conclusions

Conditional Density Estimation

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$$

MS, Takeuchi, Suzuki, Kanamori, Hachiya & Okanohara (IEICE-ED2010)

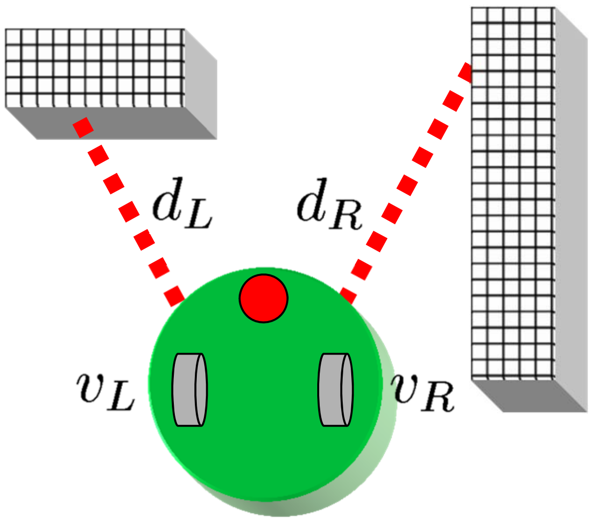
- Regression = Conditional mean estimation
- However, regression is not informative enough for complex data analysis:
  - Multi-modality
  - Asymmetry
  - Hetero-scedasticity
- Directly estimation of conditional density via density-ratio estimation.



Experiments: Transition Estimation for Mobile Robot

- Transition probability $p(s'|s,a)$: Probability of being at state s' when action a is taken at s.

Khepera robot
- State: Infrared sensors
- Action: Wheel speed



Mean (std.) test negative log-likelihood over 10 runs (smaller is better)
(red: comparable by 5% t-test)

Data	uLSIF	ε-KDE	MDN
Khepera1	1.69(0.01)	2.07(0.02)	1.90(0.36)
Khepera2	1.86(0,01)	2.10(0.01)	1.92(0.26)
Pendulum1	1.27(0.05)	2.04(0.10)	1.44(0.67)
Pendulum2	1.38(0.05)	2.07(0.10)	1.43(0.58)
Comp. Time	**1**	**0.164**	**1134**

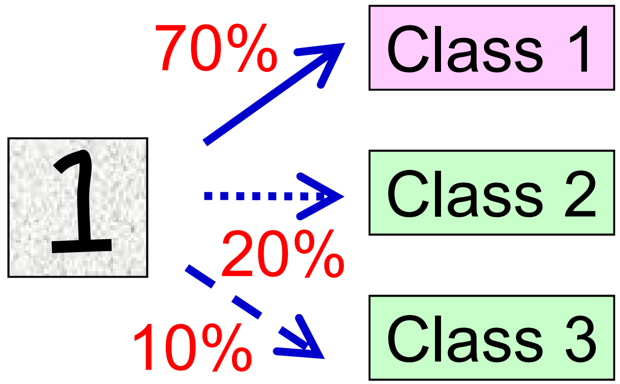
ε-KDE: ε-neighbor kernel density estimation
MDN: Mixture density network Bishop (Book2006)

Probabilistic Classification

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})}$$

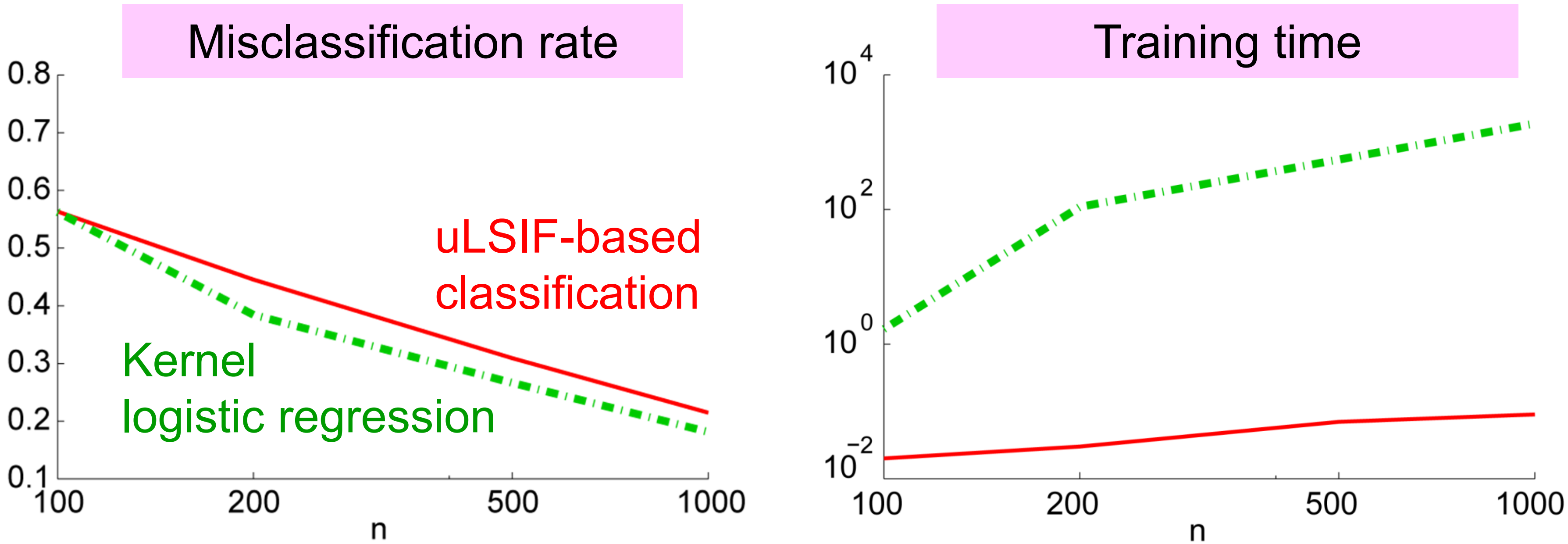
MS (IEICE-ED2010)

- If y is categorical, conditional probability estimation corresponds to learning class-posterior probability.
- Least-squares density ratio estimation (uLSIF) provides an analytic estimator:
  - Computationally efficient alternative to kernel logistic regression.
  - No normalization term included.
  - Classwise training is possible.



Numerical Example

- Letter dataset (26 classes):



- uLSIF-based classification method:
  - Comparable accuracy with KLR.
  - Training is 1000 times faster!

More Experiments

Yamada, MS, Wichern & Simm (IEICE2011)

Dataset	uLSIF	KLR
Aeroplane	82.6(1.0)	83.0(1.3)
Bicycle	77.7(1.7)	76.6(3.4)
Bird	68.7(2.0)	70.8(2.2)
Boat	74.4(2.0)	72.8(2.6)
Bottle	65.4(1.8)	62.1(4.3)
Bus	85.4(1.4)	85.6(1.4)
Car	73.0(0.8)	72.1(1.2)
Cat	73.6(1.4)	74.1(1.7)
Chair	71.0(1.0)	70.5(1.0)
Cow	71.7(3.2)	69.3(3.6)
Diningtable	75.0(1.6)	71.4(2.7)
Dog	69.6(1.0)	69.4(1.8)
Horse	64.4(2.5)	61.2(3.2)
Motorbike	77.0(1.7)	75.9(3.3)
Person	67.6(0.9)	67.0(0.8)
Pottedplant	66.2(2.6)	61.9(3.2)
Sheep	77.8(1.6)	74.0(3.8)
Sofa	67.4(2.7)	65.4(4.6
Train	79.2(1.3)	78.4(3.0)
Tvmonitor	76.7(2.2)	76.6(2.3)
Training time [sec]	**0.7**	**24.6**

Pascal VOC 2010 image classification: Mean AUC (std) over 50 runs (red: comparable by 5% t-test)

Freesound audio tagging: Mean AUC (std) over 50 runs

	uLSIF	KLR
AUC	70.1(9.6)	66.7(10.3)
Training time [sec]	0.005	0.612

Other Applications

- Action recognition from accelerometer

Hachiya, MS & Ueda (Neurocomputing 2011)

- Age prediction from faces

Ueki, MS, Ihara & Fujita (ACPR2011)

Organization of This Lecture

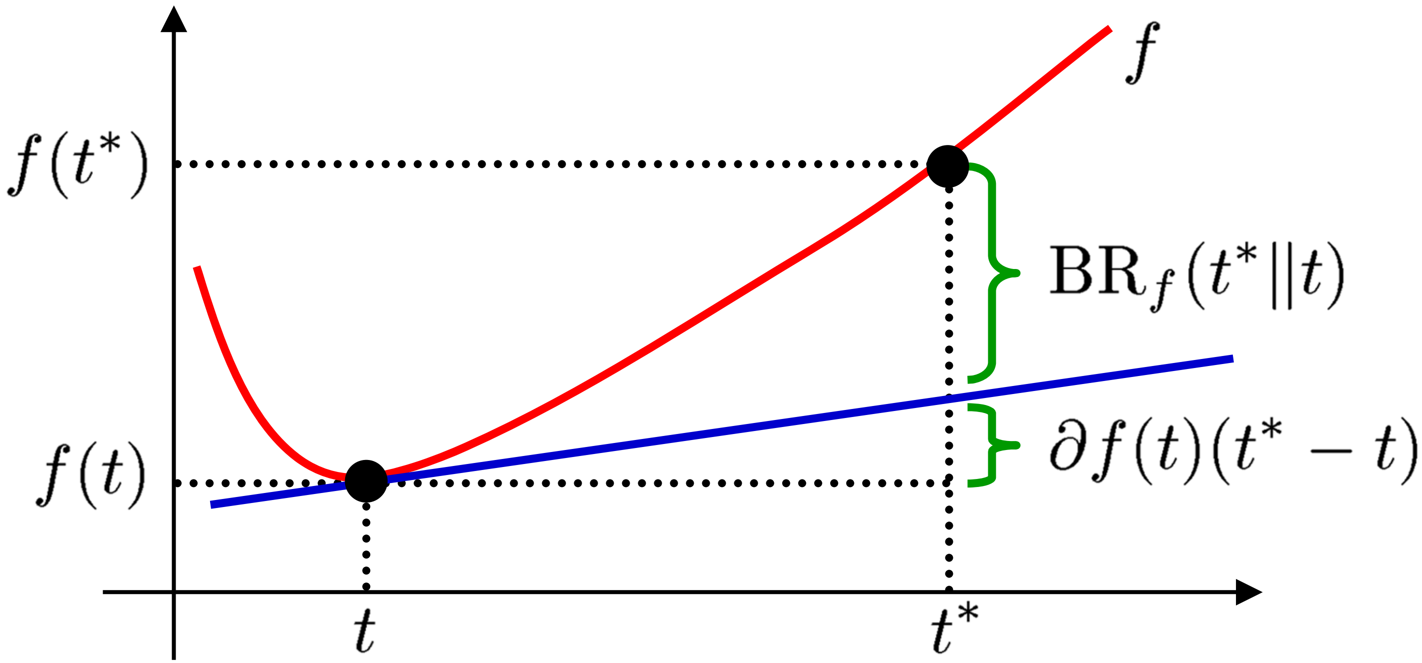
1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
4. More on Density Ratio Estimation
   A) Unified Framework
   B) Dimensionality Reduction
   C) Relative Density Ratios
5. Conclusions

Bregman (BR) Divergence

Bregman (1967)

- f : Differentiable convex function
- BR divergence with function f :

$$\mathrm{BR}_f(t^*\|t) := f(t^*) - f(t) - \underbrace{\partial f(t)(t^* - t)}_{\text{Linear prediction from } f(t) \text{ to } f(t^*)}$$



Density-Ratio Fitting under BR Divergence

MS, Suzuki & Kanamori (AISM2012)

- Fit a ratio model $\widehat{r}(\boldsymbol{x})$ to true ratio $r(\boldsymbol{x})$ under the BR divergence:

$$\min_{\widehat{r}} \mathrm{BR}_f\left(\widehat{r}\right)$$

$$\begin{aligned}\mathrm{BR}_f\left(\widehat{r}\right) &= \int p_{\mathrm{de}}(\boldsymbol{x})\nabla f(\widehat{r}(\boldsymbol{x}))\widehat{r}(\boldsymbol{x})d\boldsymbol{x} - \int p_{\mathrm{de}}(\boldsymbol{x})f(\widehat{r}(\boldsymbol{x}))d\boldsymbol{x} \\ &\quad - \int p_{\mathrm{nu}}(\boldsymbol{x})\nabla f(\widehat{r}(\boldsymbol{x}))d\boldsymbol{x} + C \\ &\approx \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\nabla f(\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}}))\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}}) - \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} f(\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}})) \\ &\quad - \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\nabla f(\widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}})) + C\end{aligned}$$

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

Unified View

- Logistic regression:

$$f(t) = t\log t - (1+t)\log(1+t)$$

- (Extended) kernel mean matching:

$$f(t) = (t-1)^2/2 \qquad \min_{\widehat{r}} \|\nabla J(\widehat{r})\|^2$$

- KL-based method:

$$f(t) = t\log t - t$$

- uLSIF:

$$f(t) = (t-1)^2/2 \qquad \min_{\widehat{r}} J(\widehat{r})$$

- Robust estimator (power divergence):

$$f(t) = \alpha^{-1}(t^{1+\alpha} - t) \quad \alpha > 0$$

Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
4. More on Density Ratio Estimation
   - A) Unified Framework
   - B) Dimensionality Reduction
   - C) Relative Density Ratios
5. Conclusions

Direct Density-Ratio Estimation with Dimensionality Reduction (D^3)

- Directly density-ratio estimation without density estimation is promising.
- However, for high-dimensional data, density-ratio estimation is still challenging.
- We combine direct density-ratio estimation with dimensionality reduction!

Hetero-distributional Subspace (HS)

MS, Kawanabe & Chui (NN2010)

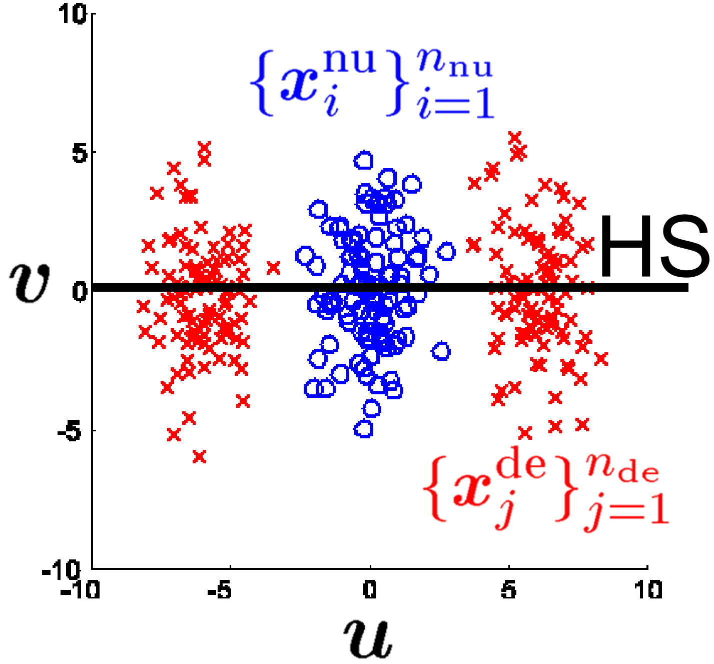
- Key assumption: $p_{\mathrm{nu}}(\boldsymbol{x})$ and $p_{\mathrm{de}}(\boldsymbol{x})$ are different only in a subspace (called HS).

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \boldsymbol{x}$$

$$= \frac{p(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{nu}}(\boldsymbol{u})}{p(\boldsymbol{v}|\boldsymbol{u})p_{\mathrm{de}}(\boldsymbol{u})} = \frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})}$$

$\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$: Full-rank and orthogonal



- This allows us to estimate the density ratio only within the low-dimensional HS!

Characterization of HS

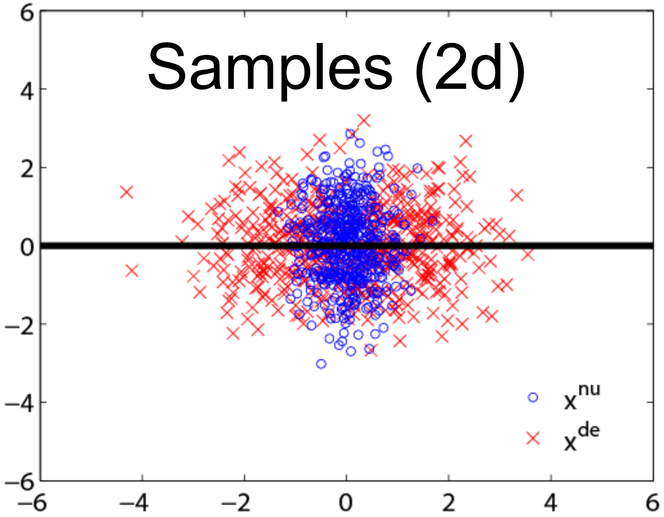
MS, Yamada, von Bünau, Suzuki, Kanamori & Kawanabe (NN2011)

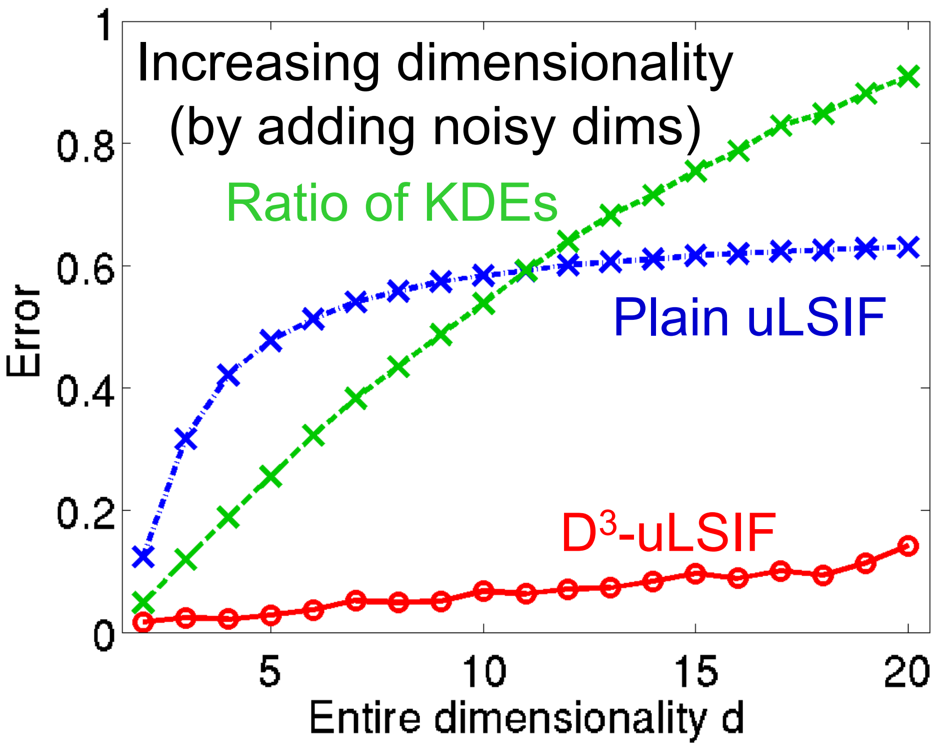
- HS is given as the maximizer of the Pearson divergence with respect to $\boldsymbol{U}$:

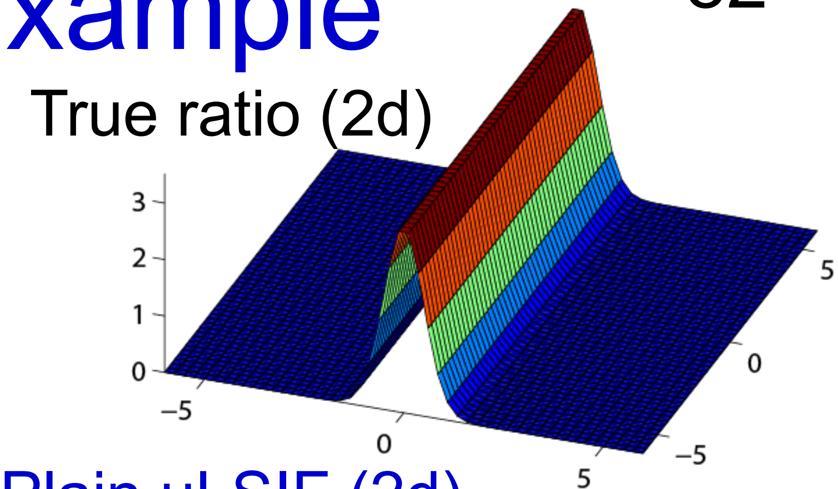
$$\mathrm{PE}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] = \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} - 1 \right)^2 p_{\mathrm{de}}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$$

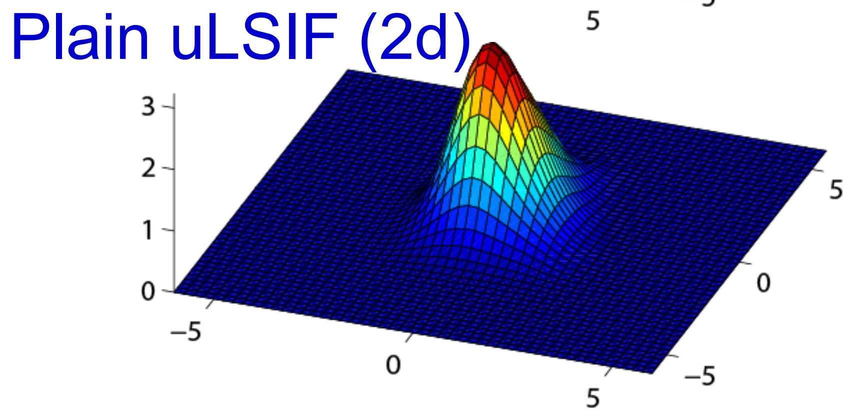
- PE can be analytically approximated by uLSIF (with good convergence property).
- HS search by
  - Natural gradient
  - A heuristic update Yamada & MS (AAAI2011)

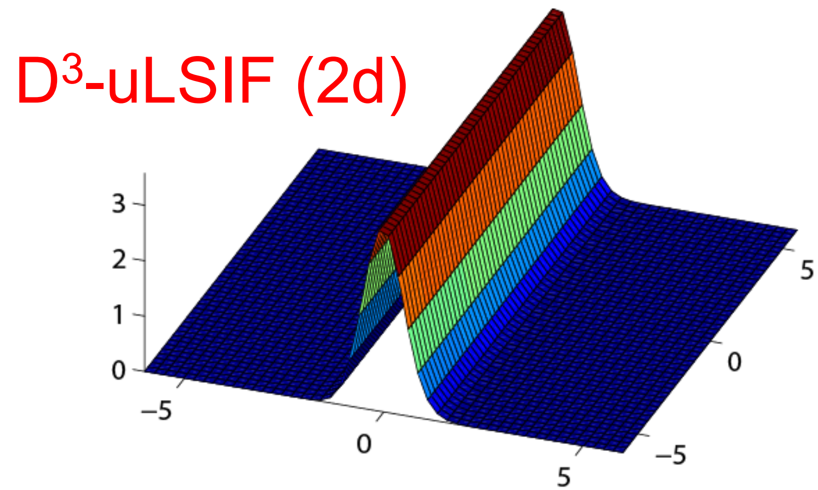
Numerical Example









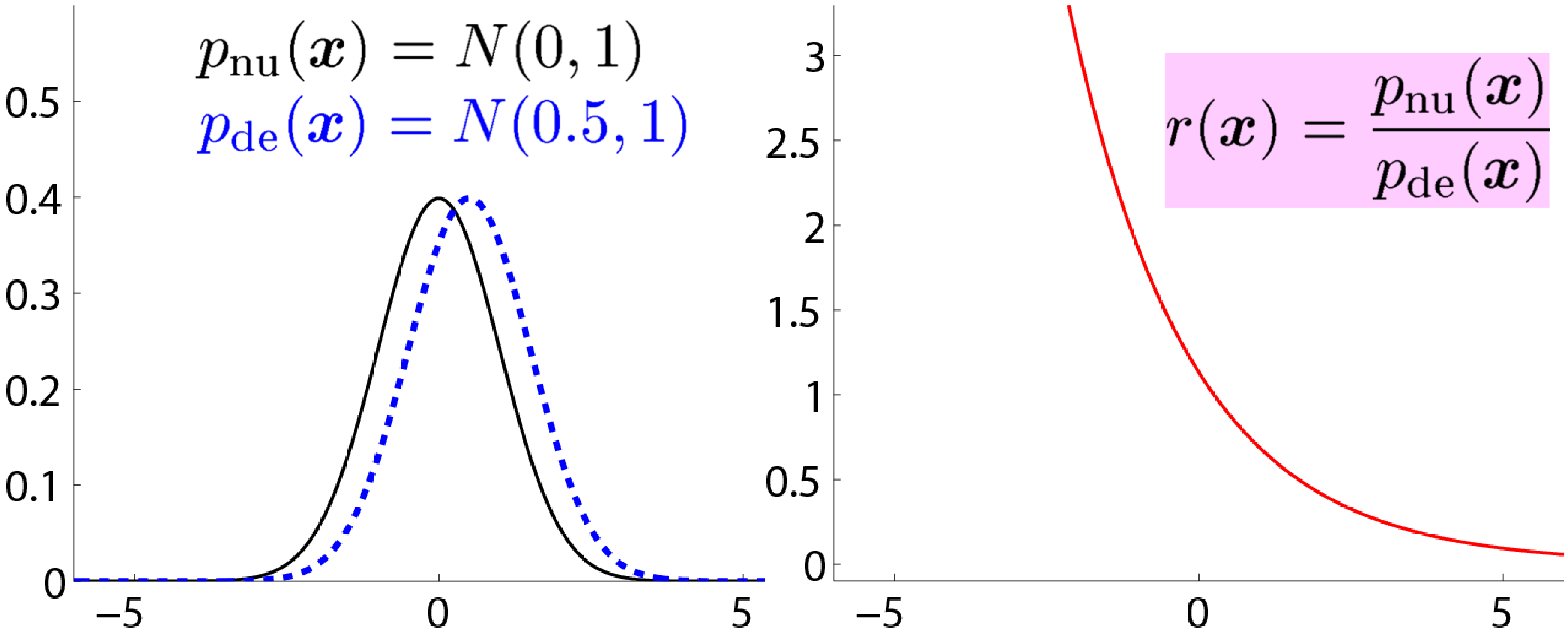


Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
4. More on Density Ratio Estimation
   - A) Unified Framework
   - B) Dimensionality Reduction
   - C) Relative Density Ratios
5. Conclusions

Weakness of Density Ratios

- Density ratio can diverge to infinity:



- Estimation becomes unreliable!

Relative Density Ratios

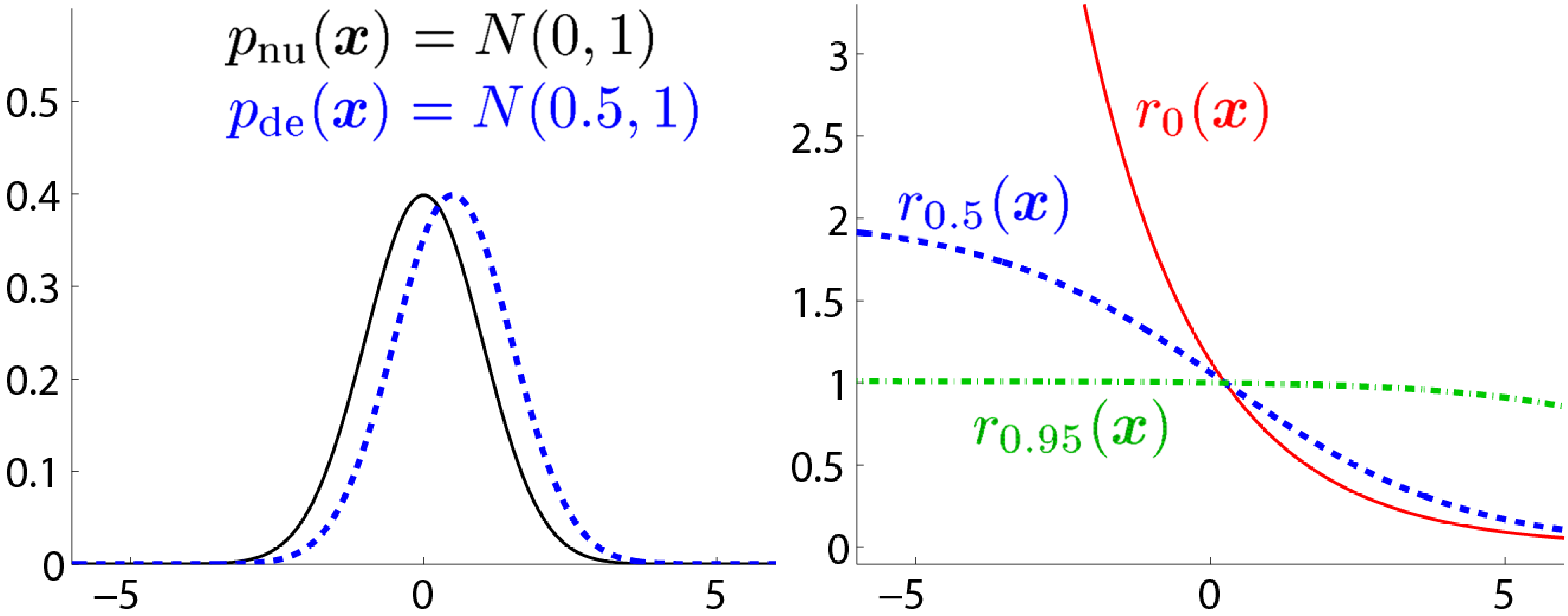
Yamada, Suzuki, Kanamori, Hachiya & MS (NIPS2011)

$$r_\beta(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{q_\beta(\boldsymbol{x})}$$

$$0 \le \beta < 1$$

$$q_\beta(\boldsymbol{x}) = \beta p_{\mathrm{nu}}(\boldsymbol{x}) + (1-\beta) p_{\mathrm{de}}(\boldsymbol{x})$$

- Bounded for any $p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})$: $r_\beta(\boldsymbol{x}) < \frac{1}{\beta}$



Estimation of Relative Ratios

- Linear model: $$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

- Relative unconstrained least-squares importance fitting (RuLSIF):

$$\min_{\widehat{r}} \int \left(\widehat{r}(\boldsymbol{x}) - r_\beta(\boldsymbol{x})\right)^2 q_\beta(\boldsymbol{x}) d\boldsymbol{x} \qquad r_\beta(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{q_\beta(\boldsymbol{x})}$$

$$q_\beta(\boldsymbol{x}) = \beta p_{\mathrm{nu}}(\boldsymbol{x}) + (1-\beta) p_{\mathrm{de}}(\boldsymbol{x})$$

- The solution can be computed analytically:

$$\mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right] = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$$

$$\widehat{\boldsymbol{H}} = \frac{\beta}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}}) \boldsymbol{\phi}(\boldsymbol{x}_j^{\mathrm{de}})^\top + \frac{1-\beta}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{\phi}(\boldsymbol{x}_i^{\mathrm{nu}}) \boldsymbol{\phi}(\boldsymbol{x}_i^{\mathrm{nu}})^\top \qquad \widehat{\boldsymbol{h}} = \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{\phi}(\boldsymbol{x}_i^{\mathrm{nu}})$$

Relative Pearson Divergence

$$\mathrm{PE}_\beta[p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})] = \frac{1}{2}\int \Big(r_\beta(\boldsymbol{x}) - 1\Big)^2 q_\beta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$r_\beta(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{q_\beta(\boldsymbol{x})} \qquad q_\beta(\boldsymbol{x}) = \beta p_{\mathrm{nu}}(\boldsymbol{x}) + (1-\beta)p_{\mathrm{de}}(\boldsymbol{x})$$

- Relative Pearson divergence can be more reliably approximated:

$$\widehat{\mathrm{PE}}_\beta - \mathrm{PE}_\beta = \mathcal{O}_p(n^{-1/2}c\|r_\beta\|_\infty + \lambda_n \max(1, R(r_\beta)^2))$$

$$n = \min(n_{\mathrm{nu}}, n_{\mathrm{de}}) \qquad \lambda_n \to o(1) \ \text{ and } \ \lambda_n^{-1} = o(n^{2/(2+\gamma)}),\ 0 < \gamma < 2$$

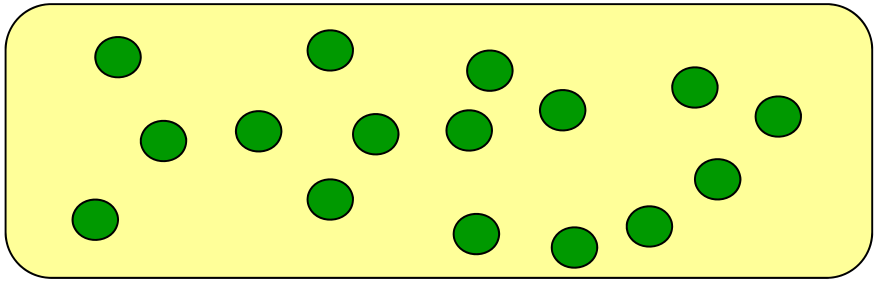
$$\|r_\beta\|_\infty = \max_{\boldsymbol{x}} r_\beta(\boldsymbol{x}) = \left\|\Big(\beta + (1-\beta)/r(\boldsymbol{x})\Big)^{-1}\right\|_\infty < \frac{1}{\beta} \qquad r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

Organization of This Lecture

1. Introduction
2. Methods of Density Ratio Estimation
3. Usage of Density Ratios
4. More on Density Ratio Estimation
5. Conclusions

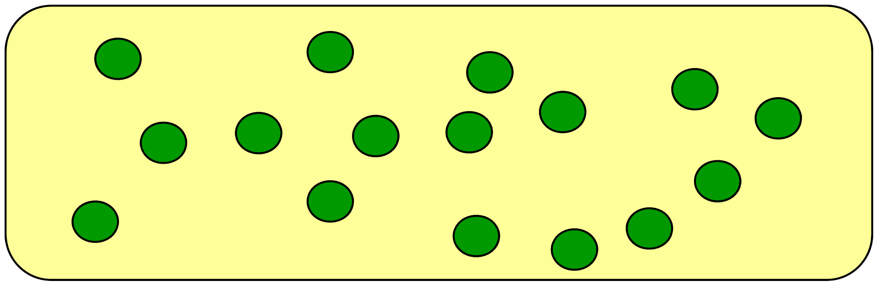
Task-Independent vs. Task-Specific

- Task-independent approach to ML:
  - Solving an ML task via the estimation of data generating distributions.
  - Applicable to solving any ML tasks.
  - No need to develop algorithms for each task.
  - However, distribution estimation is performed without regards to the task-specific goal.
  - Small error in distribution estimation can cause a big error in the target task.



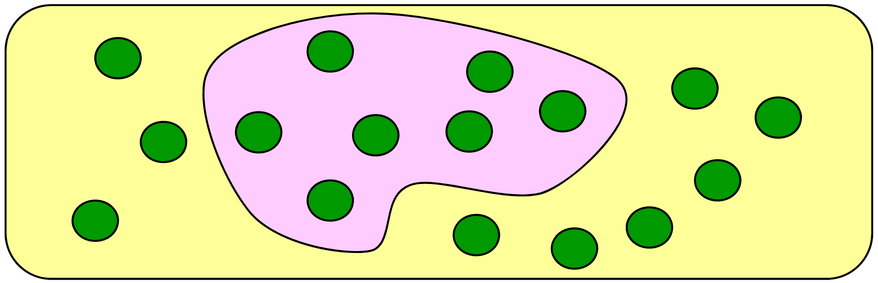
Task-Independent vs. Task-Specific

- Task-specific approach to ML:
  - Solve a target ML task directly without the estimation of data generating distributions.
  - Task-specific algorithms can be accurate.
  - However, it is cumbersome/difficult to develop tailored algorithms for every ML task.



ML for a Group of Tasks

- Density ratio estimation:
  - Develop tailored algorithms not for each task, but for a group of tasks sharing similar properties.
  - Small effort to improving the accuracy and computational efficiency contributes to enhancing the performance of many ML tasks!



- Sibling: Density difference estimation
  - Differences are more stable than ratios.

MS, Suzuki, Kanamori, Du Plessis, Liu & Takeuchi (NIPS2012)

The World of Density Ratios

Real-world applications:

Brain-computer interface, robot control, image understanding, speech recognition, natural language processing, bioinformatics

Machine learning algorithms:

- Importance sampling (covariate shift adaptation, multi-task learning)
- Distribution comparison (outlier detection, change detection in time series, two-sample test)
- Mutual information estimation (independence test, feature selection, feature extraction, clustering, independent component analysis, object matching, causal inference)
- Conditional probability estimation (conditional density estimation, probabilistic classification)

Density ratio estimation:

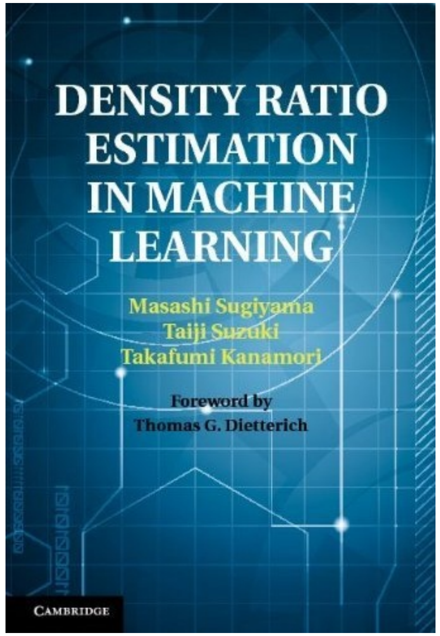
Fundamental algorithms (LogReg, KMM, KLIEP, uLSIF)
large-scale, high-dimensionality, stabilization, robustification, unification

Theoretical analysis:

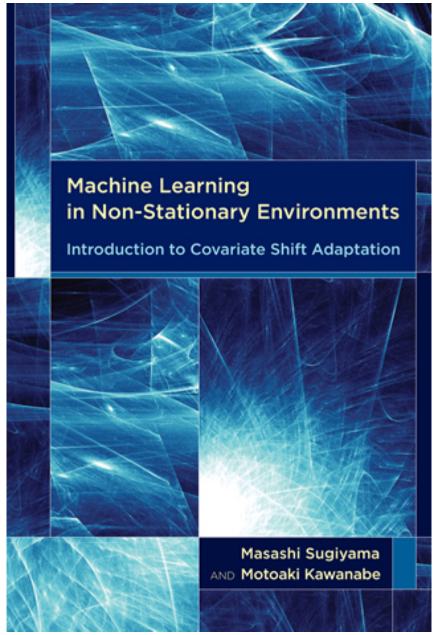
Consistency, convergence rate, information criteria, numerical stability

Books on Density Ratios

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



- Sugiyama & Kawanabe
  Machine Learning
  in Non-Stationary Environments,
  MIT Press, 2012



Acknowledgements

- Colleagues: Hirotaka Hachiya, Shohei Hido, Yasuyuki Ihara, Hisashi Kashima, Motoaki Kawanabe, Manabu Kimura, Masakazu Matsugu, Shin-ichi Nakajima, Klaus-Robert Müller, Jun Sese, Jaak Simm, Ichiro Takeuchi, Masafumi Takimoto, Yuta Tsuboi, Kazuya Ueki, Paul von Bünau, Gordon Wichern, Makoto Yamada.
- Funding Agencies: Ministry of Education, Culture, Sports, Science and Technology, Alexander von Humboldt Foundation, Okawa Foundation, Microsoft Institute for Japanese Academic Research Collaboration Collaborative Research Project, IBM Faculty Award, Mathematisches Forschungsinstitut Oberwolfach Research-in-Pairs Program, Asian Office of Aerospace Research and Development, Support Center for Advanced Telecommunications Technology Research Foundation, Japan Science and Technology Agency

- Papers and software of density ratio estimation are available from

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