# 確率分布間の距離推定：機械学習分野における最新動向

# Distance Approximation between Probability Distributions: Recent Advances in Machine Learning

杉山 将
東京工業大学 計算工学専攻

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

**概 要**

確率分布間の距離の推定は機械学習における基礎的な研究課題の一つであり，二標本検定，変化点検知，クラスバランス推定など様々な目的に応用することができる．本稿では，確率分布の推定を介さない直接距離近似法，特に，カルバック・ライブラー距離，ピアソン距離，相対ピアソン距離，$L^2$ 距離の直接近似法を概観する．

Estimation of a distance between probability distributions is one of the fundamental challenges in machine learning, because a distance estimator can be used for various purposes such as two-sample testing, change-point detection, and class-balance estimation. In this article, we review recent advances in direct distance approximation that do not involve estimation of probability distributions. More specifically, we cover direct approximators of the Kullback-Leibler distance, the Pearson distance, the relative Pearson distance, and the $L^2$-distance.

## 1 はじめに

本稿では，$\mathbb{R}^d$ 上の二つの確率分布 $P$ と $P'$ に従って独立に生成された標本集合 $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ と $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ を用いて，$P$ と $P'$ の間の距離 $D$ を推定する問題を論じる．

距離の推定量は，二つの確率分布が同一かどうかを調べる二標本検定 [26, 13]，時系列の変化点検知 [15]，クラスバランスが変化する状況下でのクラス事前確率の推定 [7]，画像中の視覚的に顕著な物体の検出 [49]，動画 [48] やツイッター [19] からのイベント検出など，様々なタスクに応用することができる．更に，同時確率分布と周辺確率分布の積との

距離の推定量は，独立性検定 [25]，特徴選択 [35, 11]，特徴抽出 [34, 41]，正準従属性分析 [14]，オブジェクト適合 [44]，独立成分分析 [33]，クラスタリング [31, 17]，因果推定 [43] など，幅広い機械学習タスクに応用する事ができる [23]．そのため，二つの確率分布間の距離を標本から精度良く推定することは，統計学，情報理論，機械学習などの分野において重要な研究課題の一つとなっており，様々な手法が開発されてきた．

確率分布 $P$ と $P'$ 間の距離 $D(P\|P')$ のナイーブな近似法は，まず標本 $\mathcal{X}$ と $\mathcal{X}'$ から確率分布 $P$ と $P'$ の推定量 $\widehat{P}_{\mathcal{X}}$ と $\widehat{P}'_{\mathcal{X}'}$ を求め，そしてそれらの推定量を距離の定義に代入することにより，距離の近似 $D(\widehat{P}_{\mathcal{X}}\|\widehat{P}'_{\mathcal{X}'})$ を計算するという二段階の方法である．しかし，第二段階の距離推定を考慮せずに第一段階の確率分布の推定を行うため，このようなナイーブな二段階推定法は必ずしも推定精度が良くない．そこで，確率分布 $P$ と $P'$ の推定を経由せず，標本 $\mathcal{X}$ と $\mathcal{X}'$ から直接的に距離の近似 $\widehat{D}(\mathcal{X}\|\mathcal{X}')$ を求めるアプローチが近年盛んに研究されるようになった [30, 20, 12, 47, 29]．

本稿では，このような距離の直接近似法の最新動向を概観する．第 2 節で機械学習に有用な距離尺度の定義を紹介し，第 3 節でそれらの直接近似法を概観する．第 4 節では距離推定量を用いた機械学習アルゴリズムを紹介し，第 5 節で結論を述べる．

# 2 距離尺度

本節では，機械学習に有用な距離尺度の定義を紹介する．数学的には，以下の四つの条件を満たす二変数関数 $d(\cdot,\cdot)$ を距離関数とよぶ：

- 非負性：$\forall x, y,\ \ d(x,y) \geq 0$
- 非退化性：$d(x,y) = 0 \iff x = y$
- 対称性：$\forall x, y,\ \ d(x,y) = d(y,x)$
- 三角不等式：$\forall x, y, z\ \ d(x,z) \leq d(x,y) + d(y,z)$

## 2.1 カルバック・ライブラー (KL) 距離

統計学や機械学習の分野で圧倒的によく用いられている距離尺度は，次式で定義される KL 距離であろう [18]：

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

ここで，$p$ と $p'$ は確率分布 $P$ と $P'$ の確率密度関数を表す．

KL 距離は，最尤推定と相性が良い，変数変換に対して不変，KL 距離がもたらすリーマン幾何構造の性質がよく調べられている [2]，直接密度比推定によって精度良く近似で

きる [30, 20, 27] など，優れた性質を持っている．しかし，KL 距離は非負性と非退化性を満たすが，対称性を持たず，三角不等式も満たさない．従って，KL 距離は厳密には距離ではない（そのため，KL ダイバージェンスとよぶこともある）．また，log 関数が含まれるため，KL 距離の近似計算には時間がかかるという問題もある．更に，log 関数が強い非線形性持つことと，密度比関数 $p/p'$ が有界でない可能性がある [4, 47] ことから，KL 距離の推定量は外れ値に弱く数値的に不安定といった弱点もある．

## 2.2 ピアソン (PE) 距離

PE 距離 [21] は KL 距離の二乗損失版であり，次式で定義される：

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

PE 距離は，観測された頻度分布が理論分布と同じかどうかを検定する適合度検定に用いられる尺度である．PE 距離と KL 距離は，共に f 距離 [1, 6] の一例となっており，変数変換に対する不変性など，理論的に似た性質を有する．

PE 距離は，KL 距離と同様に直接密度比推定によって精度良く近似できる [12, 27]．更に，PE 距離に含まれる二乗関数は最小二乗法と相性が良いことから，PE 距離は KL 距離よりも効率良く，しかも解析的に推定量を計算できるという優れた特徴を持つ．また，PE 距離は KL 距離よりも外れ値に対してロバストであることも知られている [28]．しかし，対称性を持たない，三角不等式を満たさない，密度比関数 $p/p'$ が有界でない恐れがあるといった KL 距離の問題点は，PE 距離では解決されていない．

## 2.3 相対 PE (rPE) 距離

密度比関数 $p/p'$ が有界でないという問題を解決すべく，次式で定義される rPE 距離が近年導入された [47]：

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

ここで，$q_\alpha$ は $p$ と $p'$ の $\alpha$ 混合（$0 \leq \alpha < 1$）である：

$$q_\alpha = \alpha p + (1-\alpha)p'$$

$\alpha = 0$ のとき，rPE 距離はもとの PE 距離と一致する．$p/q_\alpha$ は相対密度比とよばれ，$\alpha > 0$ のとき常に $1/\alpha$ で上から抑えられる．この相対密度比を用いることにより，rPE 距離は PE 距離の密度比の非有界性を回避している．

rPE 距離も最小二乗法と相性が良く，PE 距離とほぼ同様の方法で推定量を解析的に効率良く計算することができる．また，変数変換に対する不変性も維持されている．しか

し，対称性を持たない，三角不等式を満たさないという問題は依然として残っており，また，混合比 $\alpha$ を実際にどのように決めればよいかも明らかでない．

同様に f 距離 [1, 6] に含まれる密度比を相対密度比に置き換えることにより，より一般的な距離のクラスである相対 f 距離を考えることができる．

### 2.4 $L^2$ 距離

次式で定義される $L^2$ 距離も，標準的な距離尺度の一つである：

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

$L^2$ 距離は，非負性と非退化性に加え対称性と三角不等式も満たすため，数学的な意味での距離になっている．更に，それぞれの確率密度 $p$ と $p'$ が有界であれば，その差 $p - p'$ も常に有界である．従って，$L^2$ 距離は安定性が高く，また，rPE 距離に含まれる $\alpha$ のような調整パラメータがないため，実用上扱いやすいという利点がある．

$L^2$ 距離は最小二乗法と相性が良く，直接密度差推定 [29] により推定量を解析的に効率良く計算することができる．しかし，変数変換に対する不変性は密度比に基づく距離に対する固有の性質であるため，密度差に基づく $L^2$ 距離は変数変換に対して不変でない．

## 3 距離の直接近似

本節では，第 2 節で紹介した距離を，確率分布の推定を経由せずに直接近似する手法を概観する．

### 3.1 KL 距離の直接近似

KL 距離の直接近似の肝となる考え方は，二つの確率密度関数 $p$ と $p'$ を推定せずに密度比関数 $p/p'$ を直接推定することである [30]．具体的には，密度比モデル $r$ に関して $p$ から $r \cdot p'$ への経験一般化 KL 距離を最小にすることにより，密度比推定量 $\widehat{r}$ を求める：

$$\widehat{r} := \operatorname*{argmin}_{r} \left[\frac{1}{n'}\sum_{i'=1}^{n'} r(\boldsymbol{x}'_{i'}) - \frac{1}{n}\sum_{i=1}^{n} \log r(\boldsymbol{x}_i)\right]$$

パラメータに関して線形な密度比モデル $r(\boldsymbol{x})$ に対しては上記の最適化問題は凸となるため，勾配法などによって簡単に大域的最適解を求めることができる．例えば，ガウスカーネルモデル

$$r(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right) \tag{1}$$

はパラメータに関して線形であり，密度比が大きな値を取る領域にガウス関数を配置するため，実用上有用である [27]．ガウスカーネルのバンド幅 $\sigma$ は，上記の目的関数に対する交差確認法によって客観的に決定することができる．こうして求めた密度比推定量 $\widehat{r}$ を用いて，KL 距離推定量 $\widehat{\mathrm{KL}}(\mathcal{X}\|\mathcal{X}')$ は次式で与えられる：

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

上記の手法（KL 重要度推定法；KLIEP 法）の MATLAB® による実装は，以下のページからダウンロードできる：

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

KLIEP 法は，式 (1) の線形モデルだけでなく，対数線形モデル [39]，ガウス混合モデル [42]，確率的主成分分析器の混合モデル [46] などにも拡張されている．また，拘束条件なしのアルゴリズム [20] も独立に提案されており，これは KL 距離のルジャンドル・フェンシェル下界 [16] を近似的に最大化することに対応している：

$$\widehat{\mathrm{KL}}'(\mathcal{X}\|\mathcal{X}') := \max_{r}\left[\frac{1}{n}\sum_{i=1}^{n}\log r(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{i'=1}^{n'} r(\boldsymbol{x}'_{i'}) + 1\right]$$

## 3.2 PE 距離の直接近似

PE 距離の直接推定も，KL 距離と同様に確率密度 $p$ と $p'$ の推定を経由せずに密度比 $p/p'$ を直接推定することが肝となるが，PE 距離の推定量は KL 距離の推定量よりも効率良く計算することができる [12]．具体的には，密度比モデル $r$ と真の密度比 $p/p'$ との $p'$ による重み付き経験二乗誤差を最小にすることにより，密度比推定量 $\widehat{r}$ を求める：

$$\widehat{r} := \operatorname*{argmin}_{r}\left[\frac{1}{n'}\sum_{i'=1}^{n'} r^2(\boldsymbol{x}'_{i'}) - \frac{2}{n}\sum_{i=1}^{n} r(\boldsymbol{x}_i)\right]$$

パラメータに関して線形な密度比モデル (1) に対しては，$\ell_2$ 正則化 [10] を施した密度比推定量を解析的に求めることができ，また，一つ抜き交差確認のスコアも解析的に計算することができる [40]．一方，$\ell_1$ 正則化 [36] を施せばパラメータ $\{\theta_i\}_{i=1}^{n}$ がスパースになるため，非常に効率良く解を求めることができ [37]，また，正則化パス追跡 [9] を行うこともできる．

$\ell_2$ 正則化に対する上記の手法（拘束なし最小二乗重要度推定法;uLSIF 法）の MATLAB® による実装は，以下のページからダウンロードできる：

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

なお，密度比 $p(\boldsymbol{x})/p'(\boldsymbol{x})$ の汎関数として定義されるより一般的な f 距離 [1, 6] に対しても，直接密度比推定による距離近似が可能である [20, 30]．

### 3.3 rPE 距離の直接近似

密度 $p$ と $p'$ の $\alpha$ 混合 $q_\alpha$ に $p'$ を置き換えることにより，rPE 距離は PE 距離と同様の方法で推定することができる [47]：

$$\widehat{r} := \operatorname*{argmin}_r \left[ \frac{\alpha}{n} \sum_{i=1}^{n} r^2(\boldsymbol{x}_i) + \frac{1-\alpha}{n'} \sum_{i'=1}^{n'} r^2(\boldsymbol{x}'_{i'}) - \frac{2}{n} \sum_{i=1}^{n} r(\boldsymbol{x}_i) \right]$$

従って，rPE 距離推定は PE 距離推定と同様に優れた計算効率をもつ．

上記の手法（相対 uLSIF 法；RuLSIF 法）の MATLAB® による実装は，以下のページからダウンロードできる：

http://sugiyama-www.cs.titech.ac.jp/~yamada/RuLSIF.html

### 3.4 $L^2$ 距離の直接近似

$L^2$ 距離の直接近似の肝となる考え方は，二つの確率密度関数 $p$ と $p'$ を推定せずに密度差 $p-p'$ を直接推定することである [29]．具体的には，密度差モデル $f$ と真の密度差 $p-p'$ との経験二乗誤差を最小にすることにより，密度差推定量 $\widehat{f}$ を求める：

$$\widehat{f} := \operatorname*{argmin}_f \left[ \int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - \left( \frac{2}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - \frac{2}{n'} \sum_{i'=1}^{n'} f(\boldsymbol{x}'_{i'}) \right) \right]$$

実用上は，ガウスカーネル密度差モデル

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \exp\left( -\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2} \right) + \sum_{i'=1}^{n'} \theta_{n+i'} \exp\left( -\frac{\|\boldsymbol{x}-\boldsymbol{x}'_{i'}\|^2}{2\sigma^2} \right)$$

を用いると，目的関数の第一項目 $\int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$ が解析に計算でき都合が良い．上記の最適化問題は，PE 距離近似における最小二乗密度比推定と本質的に同じ形式のため，最小二乗密度比推定の優れた計算効率を $L^2$ 距離の近似においても享受できる．

上記の手法（最小二乗密度差推定法；LSDD 法）の MATLAB® による実装は，以下のページからダウンロードできる：

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

なお，密度差 $p(\boldsymbol{x}) - p'(\boldsymbol{x})$ の汎関数として定義されるより一般的な距離尺度に対しても，直接密度差推定による距離近似が可能である [8]．

## 4 距離推定量に基づく機械学習

本節では，距離推定量を用いた機械学習アルゴリズムを紹介する．

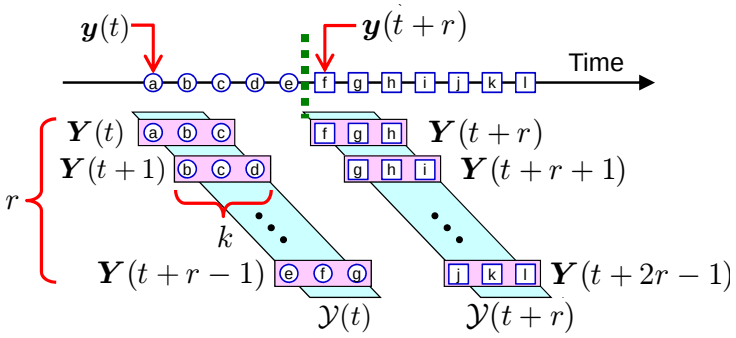


図 1: Change-point detection in time series.

## 4.1 時系列の変化点検知

時系列の背後に潜む変化を検知することが目的である．時刻 $t$ における $m$ 次元時系列標本を $\boldsymbol{y}(t) \in \mathbb{R}^m$ で表し，時刻 $t$ から長さ $k$ の部分時系列を $\boldsymbol{Y}(t) \in \mathbb{R}^{km}$ で表す：

$$\boldsymbol{Y}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top$$

ただし，$^\top$ は転置を表す．ここでは，単一の時系列標本 $\boldsymbol{y}(t)$ でなく部分系列 $\boldsymbol{Y}(t)$ を一つの標本とみなすことにより，時間に依存した情報を簡便に取り込むことにする [15]．時刻 $t$ から $r$ ステップ分の標本集合を $\mathcal{Y}(t)$ で表す：

$$\mathcal{Y}(t) := \{\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+r-1)\}$$

このとき，$\mathcal{Y}(t)$ と $\mathcal{Y}(t+r)$ の確率分布間の距離は，変化点らしさを表すスコアとして利用することができる（図 1 参照）．

これまでに，rPE 距離に基づいた変化点検知手法 [19] と $L^2$ 距離に基づいた変化点検知手法 [29] の有効性が実験的に示されている．特に，rPE 距離に基づいた変化点検知手法は，動画 [48] やツイッター [19] からのイベント検出に適用され，有望な結果が得られている．

## 4.2 クラスバランスが変化する状況下でのクラス事前確率の推定

実世界のパターン認識タスクでは，クラス間のバランスが訓練データとテストデータで異なることがある．このような場合に訓練データを用いてナイーブに分類器を学習すると，訓練データとテストデータでクラスバランスが異なるため，テストデータのラベル予測に対して強いバイアスをもたらす可能性がある．ここでは，パターン $\boldsymbol{x} \in \mathbb{R}^d$ をクラス $y \in \{+1, -1\}$ に分類する二値分類問題を考えることにし，半教師付き学習 [3] の枠組み，すなわち，ラベル付き訓練データに加えてラベルなしテストデータが与えられる状況のもとで，テストデータのクラスバランスを推定することを目標とする．

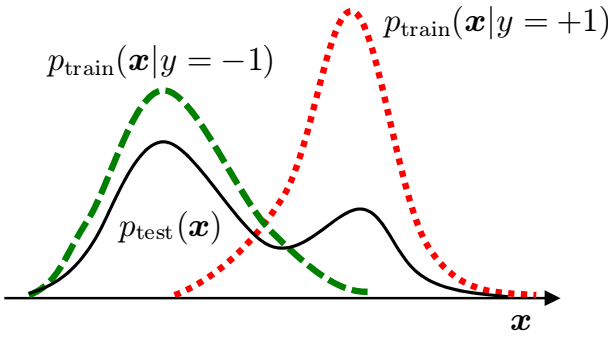


図 2: Class-prior estimation under class-balance change.

テストデータのクラスバランスは，テストパターンの確率分布に訓練パターンのクラス毎確率分布の混合分布を適合させることにより，推定することができる [7]：

$$\min_{\pi\in[0,1]} D(p_{\mathrm{test}}\|q^{\pi}_{\mathrm{test}})$$

ここで，$p_{\mathrm{test}}$ はテストパターンの確率分布を表し，$q^{\pi}_{\mathrm{test}}$ は訓練パターンのクラス毎確率分布の $\pi$ 混合を表す（図 2 参照）：

$$q^{\pi}_{\mathrm{test}}(\boldsymbol{x}) := \pi p_{\mathrm{train}}(\boldsymbol{x}|y=+1) + (1-\pi)p_{\mathrm{train}}(\boldsymbol{x}|y=-1)$$

PE 距離 [7] と $L^2$ 距離 [29] に基づいたクラスバランス推定法の有効性が，実験的に示されている．

## 4.3 画像中の視覚的に顕著な物体の検出

画像に含まれる視覚的に顕著な物体を検出することが目的である．これは，画像中のある注目領域の明るさ，エッジ，色などの画像特徴の確率分布 $p_{\mathrm{center}}$ を考え，注目領域周辺の画像特徴分布 $p_{\mathrm{surrounding}}$ との距離 $D(p_{\mathrm{center}}\|p_{\mathrm{surrounding}})$ を推定することにより，注目領域に含まれる物体の視覚的顕著度を求めることができる（図 3 参照）．すなわち，視覚的に顕著な物体がない領域では，$p_{\mathrm{center}}$ と $p_{\mathrm{surrounding}}$ がほぼ等しいため距離 $D(p_{\mathrm{center}}\|p_{\mathrm{surrounding}})$ は小さくなり，視覚的顕著度は低くなる．一方，視覚的に顕著な物体が注目領域に含まれる場合は，$p_{\mathrm{center}}$ と $p_{\mathrm{surrounding}}$ が大きく異なるため距離 $D(p_{\mathrm{center}}\|p_{\mathrm{surrounding}})$ は大きくなり，視覚的顕著度は高くなる．注目領域の場所や大きさを変えながらこの距離計算を画像全体に対して行うことにより，画像中の顕著な物体を検出することができる．

rPE 距離に基づいた視覚的顕著物体の検出アルゴリズムの有効性が，実験的に示されている [49].

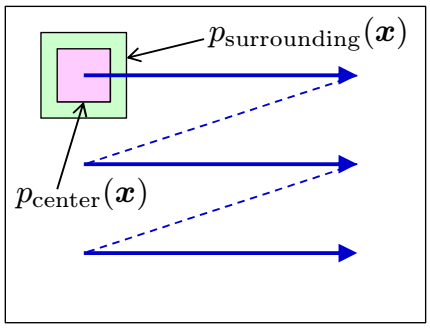


図 3: Salient-object detection in an image.

### 4.4 統計的独立性の検定

同時確率密度 $p_{\mathbf{U},\mathbf{V}}$ を持つ確率分布から独立に生成されたペア標本 $\{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^n$ を用いて，二つの確率変数 $\boldsymbol{U}$ と $\boldsymbol{V}$ の統計的従属性の強さを測ることが目的である．ここで，同時確率密度 $p_{\mathbf{U},\mathbf{V}}$ と周辺確率密度の積 $p_{\mathbf{U}} \cdot p_{\mathbf{V}}$ との間の距離を考えることにすれば，これは統計的な従属性の尺度となる．なぜならば，距離がゼロ（すなわち，$p_{\mathbf{U},\mathbf{V}} = p_{\mathbf{U}} \cdot p_{\mathbf{V}}$）の時だけ $\boldsymbol{U}$ と $\boldsymbol{V}$ は独立となり，距離が大きくなればなるほど $\boldsymbol{U}$ と $\boldsymbol{V}$ の従属性が強くなるからである．

このような従属性尺度は，距離推定アルゴリズムに与える二つのデータ集合を $\mathcal{X} = \{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^n$ と $\mathcal{X}' = \{(\boldsymbol{u}_i, \boldsymbol{v}_j)\}_{i,j=1}^n$ に変更することにより，通常の確率分布間距離と同様に推定することができる．KL 距離に基づく従属性尺度は相互情報量 [22] とよばれ，情報理論において重要な働きをする [5]．一方，PE 距離に基づく相互情報量は二乗損失相互情報量とよばれ，独立性検定 [25]，特徴選択 [35, 11]，特徴抽出 [34, 41]，正準従属性分析 [14]，オブジェクト適合 [44]，独立成分分析 [33]，クラスタリング [31, 17]，因果推定 [43] など，幅広い機械学習タスクに応用できる事が示されている [23]．$L^2$ 距離に基づく相互情報量は，二次相互情報量とよばれる [38]．

## 5 まとめ

本稿では，確率分布間の直接距離推定に関する研究の最新動向を概観した．統計学ではカルバック・ライブラー距離が圧倒的によく用いられるが，ピアソン距離，相対ピアソン距離，$L^2$ 距離の推定量は計算効率が良く，数値的に安定であり，外れ値に対して高いロバスト性を有するため，機械学習への応用ではこれらの距離推定量の方が有用であると考えられる．

本稿で紹介した直接距離推定量は，パラメトリックモデルに対して最適な $\sqrt{n}$ 一致性（$n' = n$ とする）を有することが示されており [30, 12, 47, 29], ノンパラメトリックモデルに対してミニマックス最適性を有することが証明されている [20, 30, 12, 47, 29]．また，実験的にもナイーブな密度推定に基づく距離推定量よりも精度が良いことが示されている

[30, 12, 47, 29]. しかし，直接距離推定量も次元の呪いの影響は避けられないため，高次元の距離推定問題に対しては更なる工夫が必要である．例えば，二つの確率分布が共通点を持つという期待のもと，直接距離推定に次元削減を組み合わせるという考え方 [24, 32, 45] が有望であり，更なる発展が望まれる．

著者のホームページで距離推定に関する様々な論文やソフトウェアを公開している：

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

興味を持って下さった方は，ご覧いただければ幸いである．

## 謝辞

本研究の一部は，科学技術振興機構 戦略的創造研究推進事業 個人型研究（さきがけ），最先端研究開発支援プログラム，科学研究費補助金 基盤研究（B）（一般）23300069，アジア宇宙航空研究開発事務所の支援を受けて行われた．

## 参考文献

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142, 1966.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, Providence, RI, USA, 2000.

[3] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, 2006.

[4] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 442–450, 2010.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 2006.

[6] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[7] M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In J. Langford and J. Pineau, editors, *Proceedings of 29th International Conference on Machine Learning (ICML2012)*, pages 823–830, Edinburgh, Scotland, Jun. 26–Jul. 1 2012.

[8] M. C. du Plessis and M. Sugiyama. Clustering unclustered data: Unsupervised binary labeling of two datasets having diffrent class balances. Technical Report 1305.0103, arXiv, 2013.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[10] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[11] W. Jitkrittum, H. Hachiya, and M. Sugiyama. Feature selection via $\ell_1$-penalized squared-loss mutual information. *IEICE Transactions on Information and Systems*, E95-D(7), 2013. to appear.

[12] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, Jul. 2009.

[13] T. Kanamori, T. Suzuki, and M. Sugiyama. $f$-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2012.

[14] M. Karasuyama and Sugiyama. Canonical dependency analysis based on squared-loss mutual information. *Neural Networks*, 34:46–55, 2012.

[15] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[16] A. Keziou. Dual representation of $\phi$-divergences and applications. *Comptes Rendus Mathématique*, 336(10):857–862, 2003.

[17] M. Kimura and M. Sugiyama. Dependence-maximization clustering with least-squares mutual information. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 15(7):800–805, 2011.

[18] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[19] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

[20] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[21] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[22] C. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.

[23] M. Sugiyama. Machine learning with squared-loss mutual information. *Entropy*, 15(1):80–112, 2013.

[24] M. Sugiyama, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.

[25] M. Sugiyama and T. Suzuki. Least-squares independence test. *IEICE Transactions on Information and Systems*, E94-D(6):1333–1336, 2011.

[26] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[27] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012.

[28] M. Sugiyama, T. Suzuki, and T. Kanamori. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

[29] M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 2013. to appear.

[30] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[31] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In

L. Getoor and T. Scheffer, editors, *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, pages 65–72, Bellevue, Washington, USA, Jun. 28–Jul. 2 2011.

[32] M. Sugiyama, M. Yamada, P. von Bünau, T. Suzuki, T. Kanamori, and M. Kawanabe. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2):183–198, 2011.

[33] T. Suzuki and M. Sugiyama. Least-squares independent component analysis. *Neural Computation*, 23(1):284–301, 2011.

[34] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 3(25):725–758, 2013.

[35] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52 (12 pages), 2009.

[36] R. Tibshirani. Regression shrinkage and subset selection with the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[37] R. Tomioka, T. Suzuki, and M. Sugiyama. Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation. *Journal of Machine Learning Research*, 12:1537–1586, May 2011.

[38] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[39] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

[40] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1990.

[41] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama. Computationally efficient sufficient dimension reduction via squared-loss mutual information. In C.-N. Hsu and W. S. Lee, editors, *Proceedings of the Third Asian Conference on Machine Learning (ACML2011)*, volume 20 of *JMLR Workshop and Conference Proceedings*, pages 247–262, Taoyuan, Taiwan, Nov. 13-15 2011.

[42] M. Yamada and M. Sugiyama. Direct importance estimation with Gaussian mixture models. *IEICE Transactions on Information and Systems*, E92-D(10):2159–2162, 2009.

[43] M. Yamada and M. Sugiyama. Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*, pages 643–648, Atlanta, Georgia, USA, Jul. 11–15 2010. The AAAI Press.

[44] M. Yamada and M. Sugiyama. Cross-domain object matching with model selection. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 807–815, Fort Lauderdale, Florida, USA, Apr. 11-13 2011.

[45] M. Yamada and M. Sugiyama. Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI2011)*, pages 549–554, San Francisco, California, USA, Aug. 7–11 2011. The AAAI Press.

[46] M. Yamada, M. Sugiyama, G. Wichern, and J. Simm. Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions on Information and Systems*, E93-D(10):2846–2849, 2010.

[47] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

[48] M. Yamanaka, M. Matsugu, and M. Sugiyama. Detection of activities and events without explicit categorization. *IPSJ Transactions on Mathematical Modeling and Its Applications*, 2013. to appear.

[49] M. Yamanaka, M. Matsugu, and M. Sugiyama. Salient object detection based on direct density-ratio estimation. *IPSJ Transactions on Mathematical Modeling and Its Applications*, 2013. to appear.