確率密度比を用いた新しい機械学習アルゴリズム

東京工業大学　計算工学専攻
杉山　将
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

概要

統計的機械学習のほとんどの課題は，データの生成確率分布の推定を介して解決することができる．しかし，確率分布の推定は機械学習における最も困難な問題の一つとして知られているため，現実的には分布推定を回避しながら対象となる課題を解決することが望ましい．例えば，パターン認識法の一つであるサポートベクトルマシンでは，データの生成確率を推定することなくパターン認識に最低限必要な決定境界を直接学習することにより，高い汎化性能の獲得を可能にした．

これまで，分布推定を介さない学習アルゴリズムの開発は，個々の機械学習課題に対して個別に行われてきた．それに対し我々は，様々な機械学習課題に対して統一的に分布推定を避けることのできる汎用的な枠組みを提案してきた．この枠組みでは，確率分布でなく確率密度関数の比を考え，分子と分母の確率密度関数を推定することなくそれらの比を直接推定する．密度比推定によって，重点サンプリング（共変量シフト適応，ドメイン適応，マルチタスク学習），ダイバージェンス推定（二標本検定，外れ値検出，変化点検知），相互情報量推定（独立性検定，独立成分分析，特徴選択，十分次元削減，因果推定），条件付き確率推定（確率的パターン認識，条件付き密度推定）など様々な機械学習課題が解決できることを示してきた．従って，我々の開発してきた密度比直接推定手法を用いることにより，上記の機械学習課題群を一挙に，かつ，高精度に解決できるようになった．

本講演では，この枠組の全体像を紹介すると共に，最新の研究成果についても述べる．

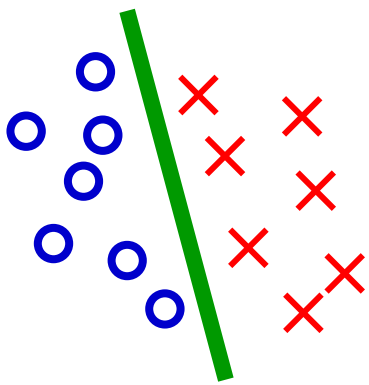
機械学習

- 機械学習：データの背後に潜む知識を自動的に獲得する
- 応用分野：音声，画像，自然言語，ロボット，脳波，生物，化学，金融など
- 具体的なタスク：パターン認識，予測，異常検出，因果推論など

- データを生成する規則（確率分布）を推定すれば，あらゆる機械学習タスクが解決できる！
  - データの生成規則を知る　＝　データの全てを知る

確率分布推定と機械学習

- しかし，確率分布の推定は，統計的データ処理において最も難しい問題の一つとされている
- 従って，ひとたび確率分布の推定を行うと，推定精度が大きく低下してしまう
- 確率分布を推定することなく，解きたいタスクを解決することが望ましい
- 例：サポートベクトルマシン（パターン認識器）
  - データの生成確率分布を推定せず，パターン認識に必要な，クラス間の分離境界だけを学習する



確率分布推定と機械学習

- 機械学習分野には様々なタスクがある:
  - 非定常環境下での適応学習, ドメイン適応, マルチタスク学習
  - 二標本検定, 異常値検出, 時系列の変化点検知,
  - 相互情報量推定, 独立性検定, 特徴選択, 十分次元削減, 独立成分分析, 因果推論,
  - 条件付き確率推定, 確率的パターン認識
- 各々のタスクに対して, 確率分布の推定を避けることのできるアルゴリズムを個別に開発することは困難

我々の提案するアプローチ：確率密度比に基づく機械学習

- 前頁の機械学習タスクは複数の確率分布を含む
- 確率密度関数の比を推定することにする
  - その際，確率密度は推定しない
  - 密度を求めるよりも，密度比を求めるほうが易しい！

$p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})$

が分かる

$\Rightarrow$

$\not\Leftarrow$

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

が分かる

- 密度比推定により，前頁の機械学習タスクが全て解決できる！

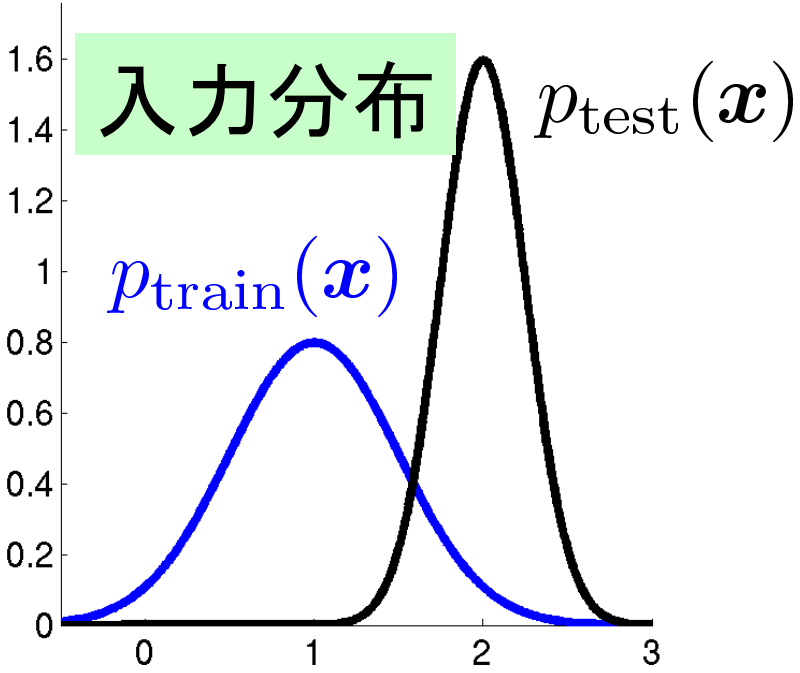
発表の流れ

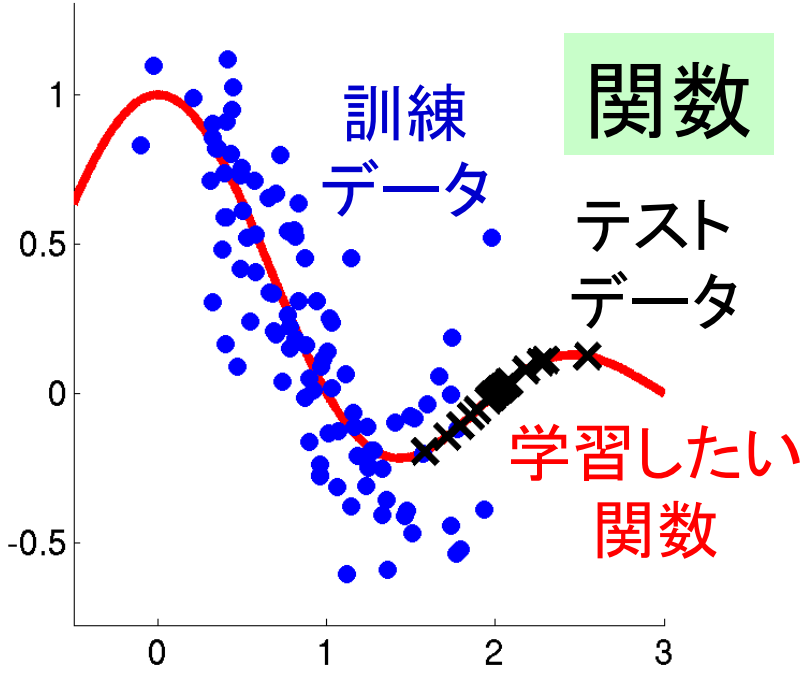
1. 密度比推定の応用例
   - A) 共変量シフト適応
   - B) 外れ値検出
   - C) 相互情報量推定
   - D) 条件付き確率推定
2. 密度比推定手法

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

共変量シフト適応

- 共変量とは入力変数の別名
- 共変量シフト: 訓練時とテスト時で入力分布が変化するが，入出力関数は変わらない
- 外挿問題が典型的な例.

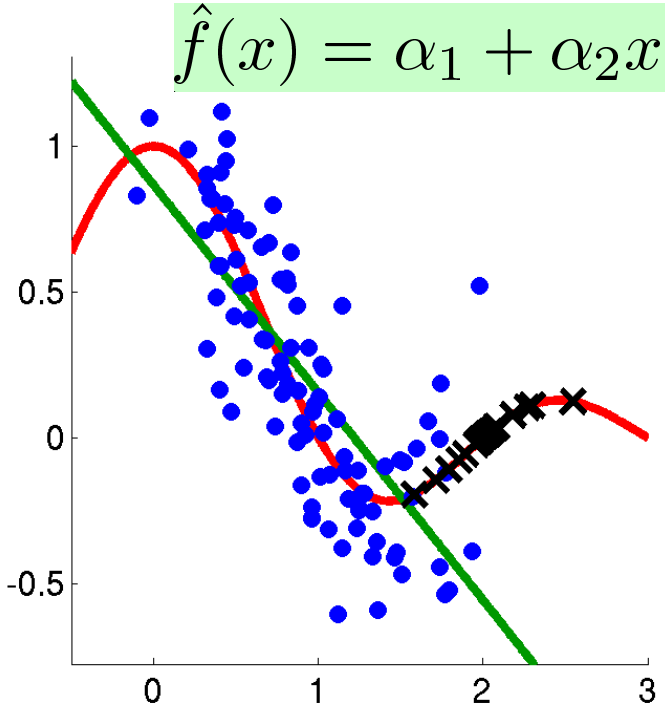




最小二乗法

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

- 通常の設定では，最小二乗法は一致性を持つ．つまり，$n \to \infty$ で最適な解が求まる
- しかし共変量シフト下では，最小二乗法は一致性を持たない



最小二乗法の一致性：大数の法則

- 標本平均は真の期待値に収束：

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

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

- 共変量シフトがなければ，これで一致性が保証される
- 訓練データ $\{\boldsymbol{x}_i\}_{i=1}^{n}$ を用いて，テストデータに対する期待値を最小にしたい！

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

重点サンプリング

- 重要度:テスト入力と訓練入力の密度の比

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

- 重要度重み付き平均:

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

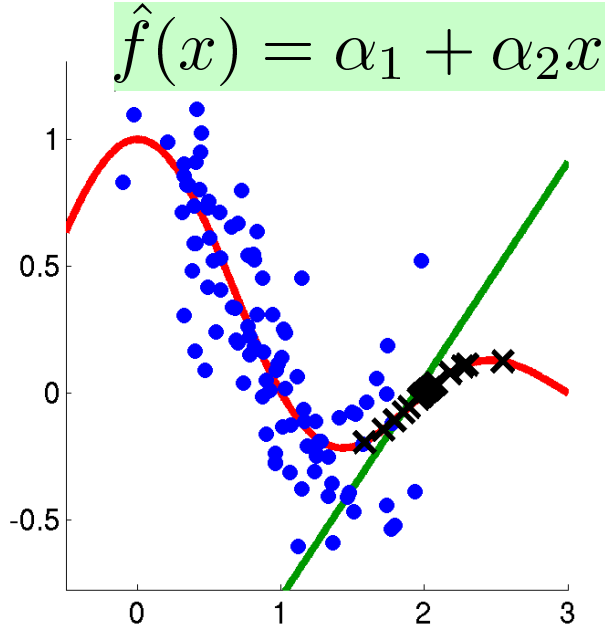
$$\longrightarrow \int \frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}\mathrm{loss}(\boldsymbol{x})p_{train}(\boldsymbol{x})d\boldsymbol{x}$$

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

重要度重み付き最小二乗法

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

- 重要度重み付き最小二乗法は共変量シフト下でも一致性を持つ
- 重要度重み付けは，あらゆる学習法に適用できる:
  - サポートベクトルマシン
  - ロジスティック回帰
  - 条件付き確率場など



“平坦化”密度比重み付け学習

Shimodaira (JSPI2000)

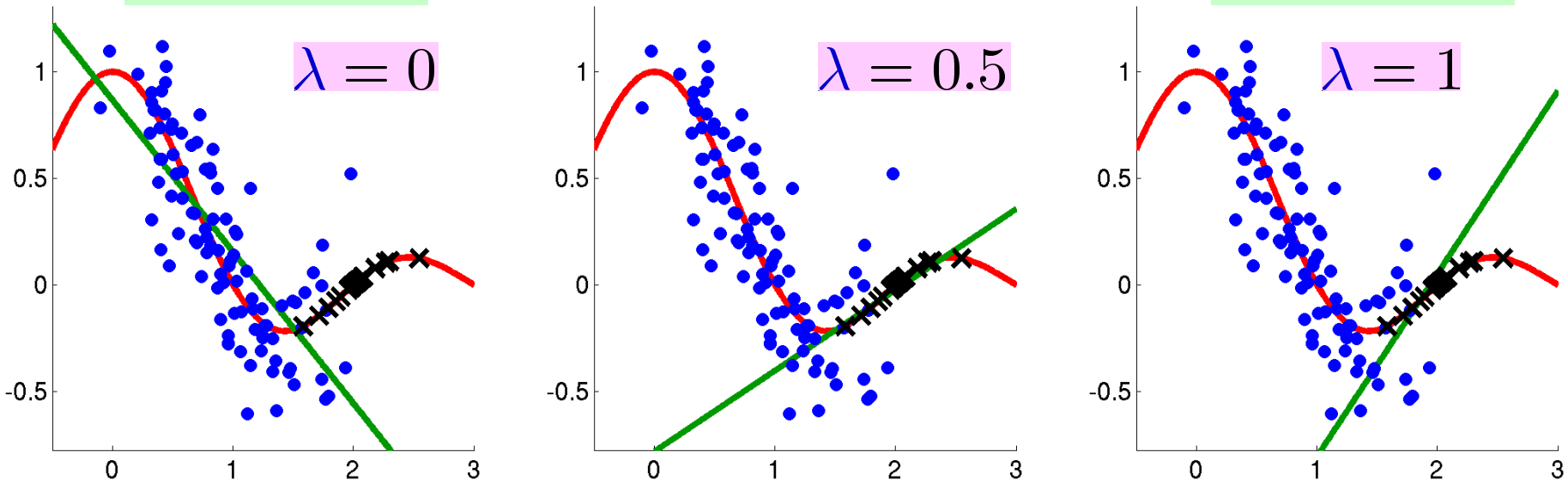
$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \right)^{\lambda} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$0 \leq \lambda \leq 1$$

バイアス：大
バリアンス：小

中間

バイアス：小
バリアンス：大



モデル選択が必要！

モデル選択

- 共変量シフト下では，通常のモデル選択規準は強いバイアスを持つ
- 密度比重み付けは，モデル選択規準に不偏性を持たせるためにも用いられる:
  - Akaike information criterion（正則モデル）

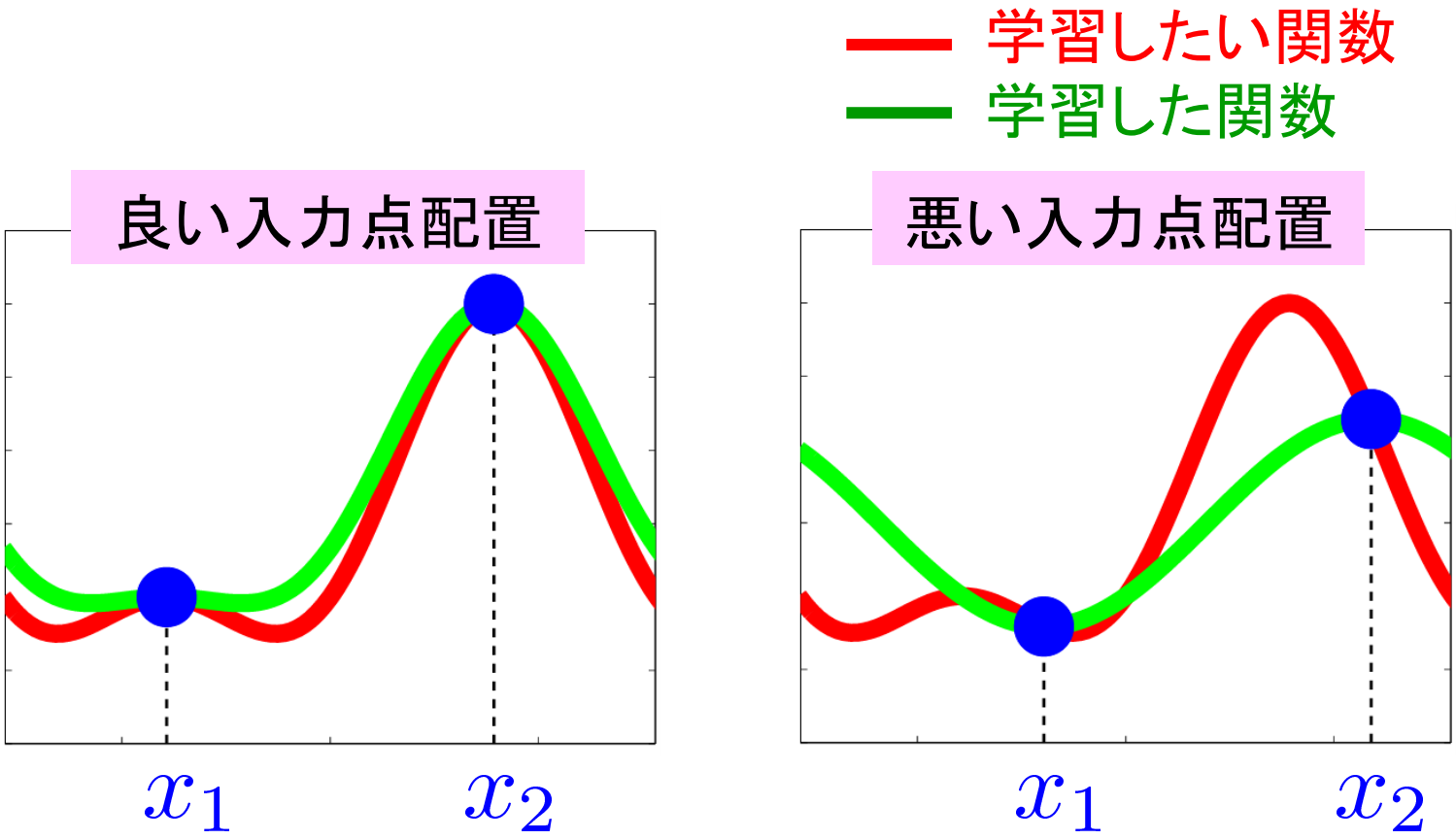
    Shimodaira (JSPI2000)
  - Subspace information criterion（線形モデル）

    Sugiyama & Müller (Stat&Dec.2005)
  - Cross-validation（任意のモデル）

    Sugiyama, Krauledat & Müller (JMLR2007)

能動学習

- 入力点の選び方は汎化能力に大きく影響する
- 目標：最適な入力点の場所を選ぶ



能動学習と共変量シフト

- 能動学習では共変量シフトが自然に起こる:
  - $p_{\mathrm{train}}(\boldsymbol{x})$ はユーザが設定
  - $p_{\mathrm{test}}(\boldsymbol{x})$ は環境から自動的に決まる
- 密度比重み付けは，能動学習のバイアスを軽減するのに役立つ:
  - 分布の基づく能動学習法: Wiens (JSPI2000)
    Kanamori & Shimodaira (JSPI2003)
    Sugiyama (JMLR2006)
  - ラベル無し標本に基づく能動学習法:
    Kanamori (NeuroComp2007)
    Sugiyama & Nakajima (MLJ2009)

実世界応用例(1)

- 顔画像からの年齢予測:
  - 照明や顔角度の変化

Ueki, Sugiyama & Ihara (ICPR2010)

- 話者識別:
  - 声質の変化

Yamada, Sugiyama & Matsui (SigPro2010)

- ブレイン・コンピュータインターフェース:
  - 心理状況の変化

Sugiyama, Krauledat & Müller (JMLR2007)
Li, Kambara, Koike & Sugiyama (IEEE-TBE2010)

実世界応用例(2)

■ 日本語のテキスト分割:

- 一般会話から医学用語へのドメイン適応

こんな／失敗／は／ご／愛敬／だ／よ／ .

Tsuboi, Kashima, Hido, Bickel & Sugiyama (JIP2008)

■ 半導体ウェハー位置合せ:

- 最適なマーカーの選択

Sugiyama & Nakajima (MLJ2009)

■ ロボット制御:

- 標本の再利用

Hachiya, Akiyama, Sugiyama & Peters (NN2009)
Hachiya, Peters & Sugiyama (ECML2009)

- 最適な探索政策の選択

Akiyama, Hachiya, & Sugiyama (NN2010)

発表の流れ

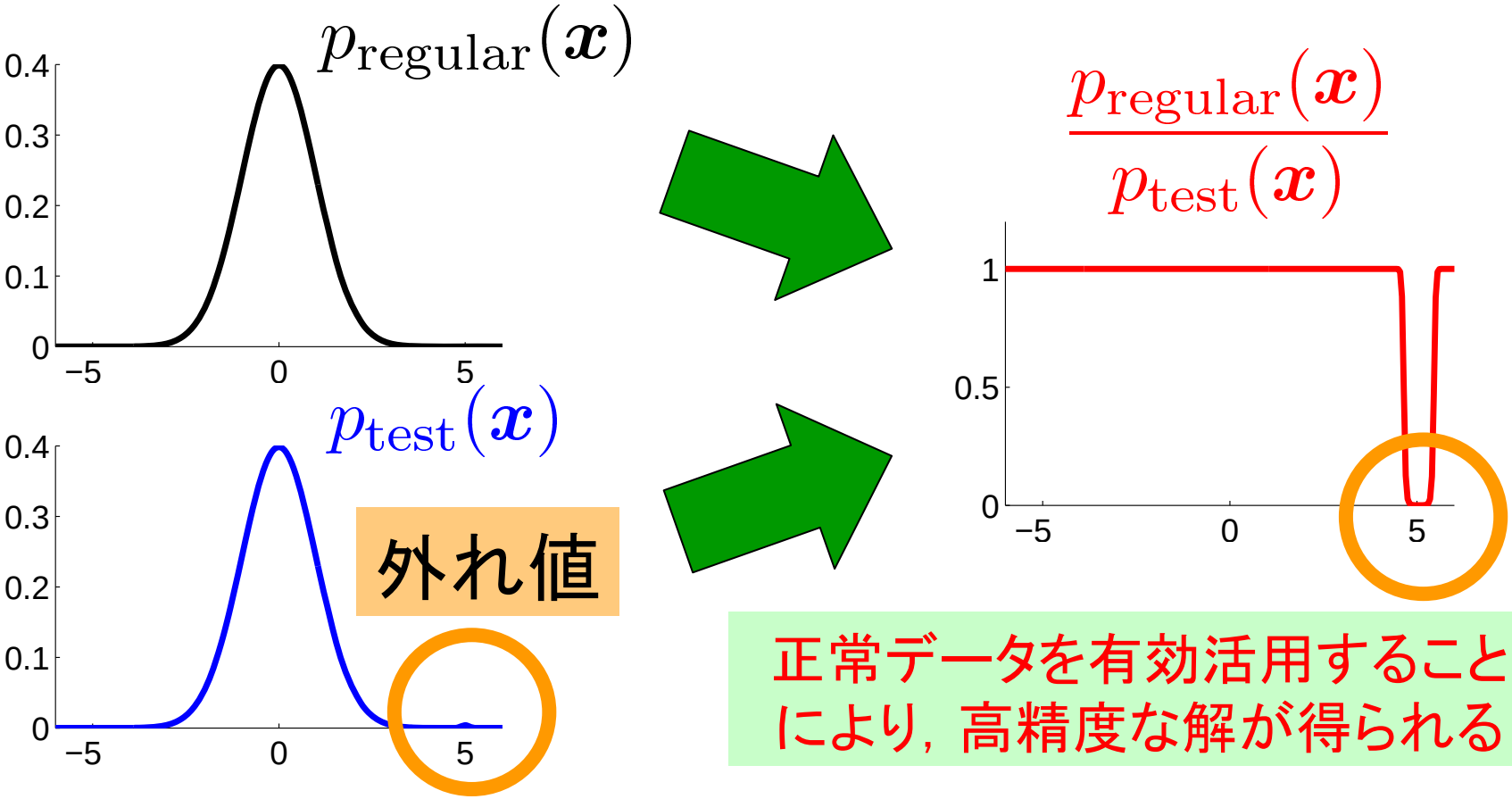
1. 密度比推定の応用例
   - A) 共変量シフト適応
   - B) 外れ値検出
   - C) 相互情報量推定
   - D) 条件付き確率推定
2. 密度比推定手法

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

正常値に基づく異常値検出

Hido, Tsuboi, Kashima, Sugiyama & Kanamori (ICDM2008, KAIS2010)
Smola, Song & Teo (AISTATS2009)

- 正常データと傾向が異なるテストデータを異常値とみなす.



正常データを有効活用することにより，高精度な解が得られる

利点

- モデル選択が可能:
  - 異常検出アルゴリズムには，正則化係数など，何らかのチューニングパラメータが含まれる
  - 異常値検出のような教師なし学習では，それらの値はヒューリスティックに決めることが多い
  - 密度比を用いたアプローチでは，密度比の推定精度の意味でモデル選択を行うことができる（例:クロスバリデーション）
- 大規模・高次元データにも適用可能

実世界応用例

- 製鉄プロセスの異常診断
- 光学部品の品質検査
  Takimoto, Matsugu & Sugiyama (DMSS2009)
- ローン顧客の審査
  Hido, Tsuboi, Kashima, Sugiyama & Kanamori (KAIS2010)
- 生体データからの睡眠診断
  Kawahara & Sugiyama (SDM2009)

二標本検定

Sugiyama, Suzuki, Ito, Kanamori & Kimura (IBISML2010-9)

- 目的：二つのデータセットの背後の確率分布が同じかどうかを検定する

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} P \qquad \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{i.i.d.}{\sim} P'$$

- アプローチ：密度比を用いて分布間の距離を推定する
  - カルバックライブラー距離：$\displaystyle\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$
  - ピアソン距離：$\displaystyle\int p'(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$

二標本検定の使用例

- 非定常環境適応：過去と現在のデータの分布を比較し（例：脳波データ），適応学習を行う必要があるかどうかを判定する
- マルチタスク学習：複数の学習タスクから得たデータの分布を比較し，似たタスクを同時に解く事により学習精度を向上させる
- データセットの統合：複数の研究室で同じ実験対象からデータを得たとき（例：マイクロアレイ），それらを統合して扱ってよいかどうかを判定する
- データベースのマッチング：複数のデータベースの属性分布を比較し，対応関係を自動認識する

発表の流れ

1. 密度比推定の応用例
   - A) 共変量シフト適応
   - B) 外れ値検出
   - C) 相互情報量推定
   - D) 条件付き確率推定
2. 密度比推定手法

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

相互情報量推定

■ 相互情報量: $$\mathrm{MI} = \int p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}$$

$\mathrm{MI} = 0$ ⟺ $\boldsymbol{x}$ と $\boldsymbol{y}$ は独立

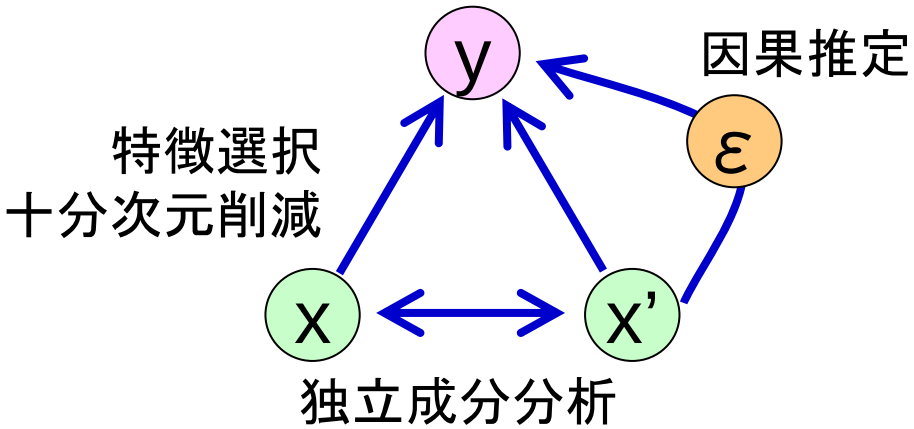
■ 相互情報量は密度比を用いて計算できる

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

Suzuki, Sugiyama, Sese & Kanamori (FSDM2008)
Suzuki, Sugiyama & Tanaka (ISIT2009)

相互情報量推定量の応用例

- 入出力間の独立性判定:
  - 特徴選択 Suzuki, Sugiyama, Sese & Kanamori (BMC Bioinformatics 2009)
  - 十分次元削減 Suzuki & Sugiyama (AISTATS2010)
- 入力間の独立性判定:
  - 独立成分分析 Suzuki & Sugiyama (NeuralComp, to appear)
- 入力と残差との独立性判定:
  - 因果推定 Yamada & Sugiyama (AAAI2010)



発表の流れ

1. 密度比推定の応用例
   - A) 共変量シフト適応
   - B) 外れ値検出
   - C) 相互情報量推定
   - D) 条件付き確率推定
2. 密度比推定手法

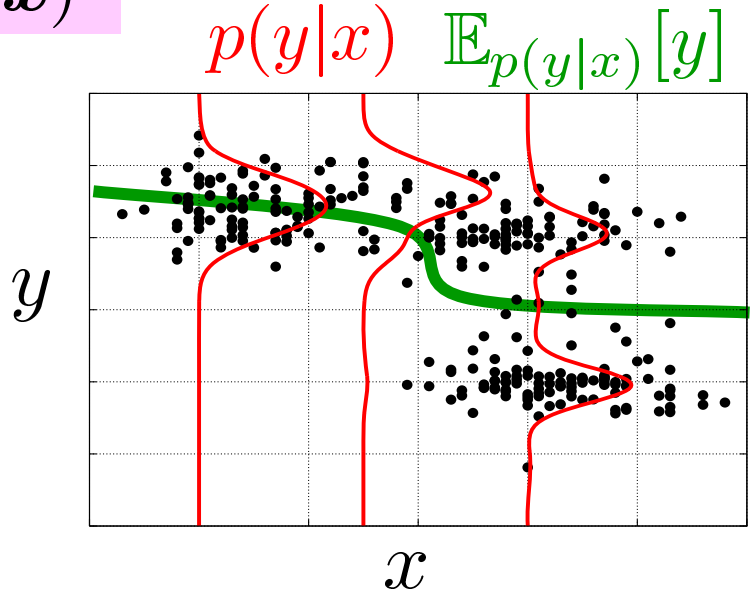
$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

条件付き確率推定

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

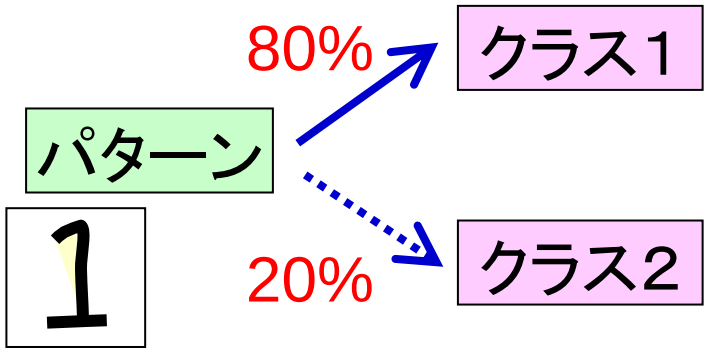
- 出力 y が連続のとき:
  - 条件付き密度推定
  - 移動ロボットの状態遷移推定

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



- 出力 y がカテゴリ値のとき:
  - 確率的分類

Sugiyama (IEICE-ED2010)



発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

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

密度比推定

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

- 密度比をデータから推定する

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

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

- ナイーブな方法：分母と分子の密度をそれぞれ推定し，その比を取る

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

Vapnikの原理

ある問題を解くとき，その問題よりも
難しい問題を途中段階で解いてはならない

$p_{\mathrm{nu}}(\boldsymbol{x}), p_{\mathrm{de}}(\boldsymbol{x})$

が分かる

$\Rightarrow$

$\not\Leftarrow$

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

が分かる

- 密度を求めるよりも，密度比を求めるほうが易しい！
- 密度推定を経由せず，密度比を推定することにする

発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

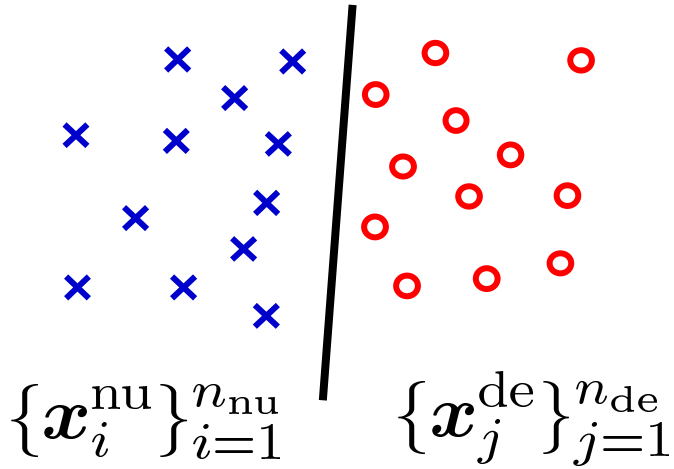
$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

確率的分類

■ 分母と分子の標本を確率的に分類する:

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

$$= \frac{p(\text{‘nu’})}{p(\text{‘de’})} \frac{p(\text{‘de’}|\boldsymbol{x})}{p(\text{‘nu’}|\boldsymbol{x})}.$$



■ ロジスティック回帰:

- モデルが正しいとき，漸近分散を最小にする

  Qin (Biometrika1998)

- モデルが正しくないときは，密度比適合法の方が良い

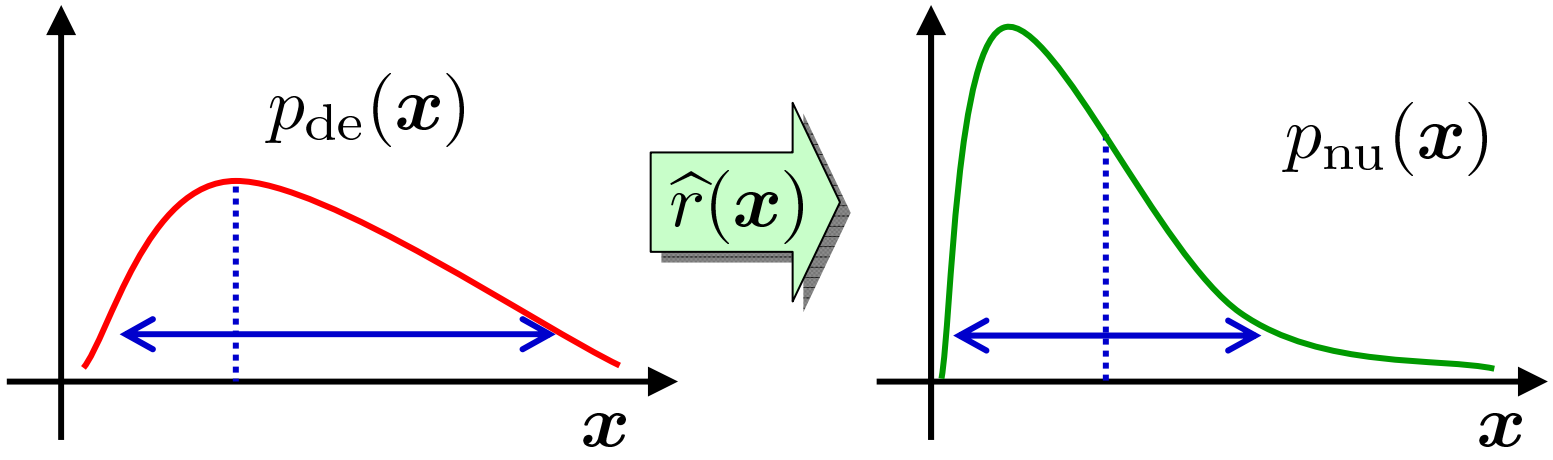
  Kanamori, Suzuki & Sugiyama (IEICE-EA2010)

発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

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

積率適合法



- $p_{\mathrm{nu}}(\boldsymbol{x})$と $\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x})p_{\mathrm{de}}(\boldsymbol{x})$ の積率を適合する
  - 例：平均（1次の積率）の適合

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

$$\min_{\widehat{r}} \left\| \int \boldsymbol{x}\widehat{r}(\boldsymbol{x})p_{\mathrm{de}}(\boldsymbol{x})d\boldsymbol{x} - \int \boldsymbol{x}p_{\mathrm{nu}}(\boldsymbol{x})d\boldsymbol{x} \right\|^2$$

カーネル平均適合

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

- 有限個の積率を適合しても真の密度比が得られるとは限らない
- カーネル平均適合:
  - ガウス再生核ヒルベルト空間内で平均を適合することにより，全ての次数の積率を効率よく適合できる

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

  - 二次計画問題の解として，$\{r(\boldsymbol{x}_j^{\mathrm{de}})\}_{j=1}^{n_{\mathrm{de}}}$ の推定値が得られる
  - 密度比関数全体の推定や，様々な損失関数にも拡張できる

Kanamori, Suzuki & Sugiyama (arXiv2009)

発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

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

密度適合法

Sugiyama, Nakajima, Kashima, von Bünau & Kawanabe (NIPS2007)

- $p_{\mathrm{nu}}(\boldsymbol{x})$ から $\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x})$ への
カルバック・ライブラー（KL）距離を最小にする

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

- KL距離の分解：

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

定式化

- 目的関数:

$$\min_{\widehat{r}} \mathrm{KL}[p_{\mathrm{nu}}(\boldsymbol{x}) || \widehat{p}_{\mathrm{nu}}(\boldsymbol{x})] \Leftrightarrow \max_{\widehat{r}} \int p_{\mathrm{nu}}(\boldsymbol{x}) \log \widehat{r}(\boldsymbol{x}) d\boldsymbol{x}$$

- 拘束条件: $\widehat{p}_{\mathrm{nu}}(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x})$ は確率密度

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

- 線形密度比モデルを用いる:

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

$$\phi_\ell(\boldsymbol{x}) \geq 0$$

（例:ガウスカーネル）

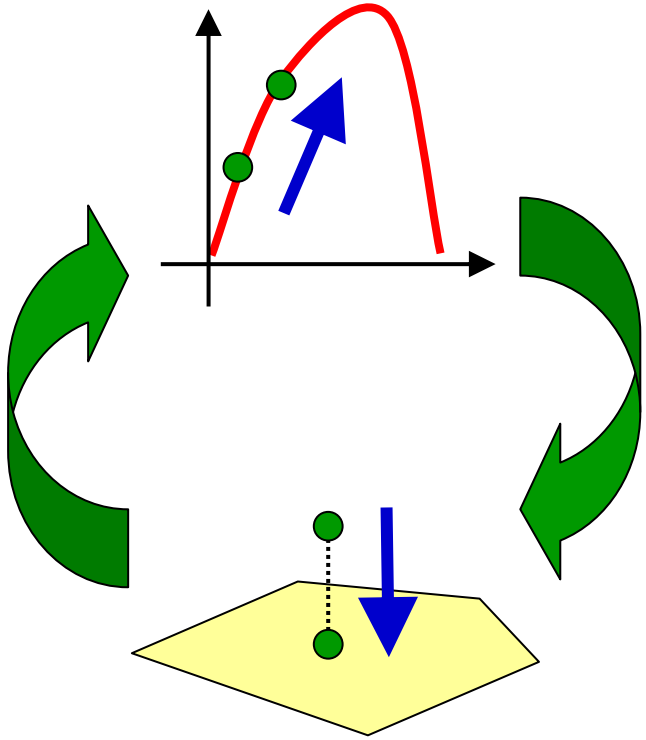
カルバック・ライブラー重要度推定法(KLIEP)

- 期待値を標本平均で近似する:

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

- 凸最適化問題なので,
  - 勾配上昇
  - 制約充足(射影)

  の繰り返しで大域的最適解を求められる
- 大域的最適解はスパース



理論解析

Nguyen, Wainwright & Jordan (NIPS2007)
Sugiyama, Suzuki, Nakajima, Kashima, von Bünau & Kawanabe (AISM2008)

■ パラメトリックモデルの場合 : $\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x})$

- 学習したパラメータは $1/\sqrt{n}$ の速さで最適値に収束
- 最適な収束率を達成している

$n = \min(n_{\mathrm{nu}}, n_{\mathrm{de}})$

■ ノンパラメトリックモデルの場合 : $\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{n_{\mathrm{nu}}} \alpha_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell^{\mathrm{nu}})$

- 学習した関数は $1/\sqrt{n}$ より少し遅い速さで真の関数に収束（被覆数やブラケットエントロピーに依存）
- 最適な収束率を達成している

拡張

- KLIEPは，線形モデル以外に
  - 線形対数モデル

    Tsuboi, Kashima, Hido, Bickel & Sugiyama (JIP2009)
  - ガウス混合モデル

    Yamada & Sugiyama (IEICE-ED2009)
  - 確率的主成分分析(PCA)混合モデル

    Yamada, Sugiyama, Wichern & Jaak (IEICE-ED2010)

  にも拡張できる

発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

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

最小二乗密度比適合(LSIF)

Kanamori, Hido & Sugiyama
(NIPS2008, JMLR2009)

- 真の密度比 $r(\boldsymbol{x})$ とそのモデル $\widehat{r}(\boldsymbol{x})$ の二乗誤差を最小にするように学習:

$$\min_{\widehat{r}} \left[J(\widehat{r}) \right]$$

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

$$J(\widehat{r}) = \frac{1}{2} \int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x}) \right)^2 p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \frac{1}{2} \int \left(\widehat{r}(\boldsymbol{x}) \right)^2 p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x} - \int \widehat{r}(\boldsymbol{x}) p_{\mathrm{nu}}(\boldsymbol{x}) d\boldsymbol{x} + C$$

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

アルゴリズム

■ 線形密度比モデルを用いる:

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

- cLSIF: 非負拘束あり, ℓ_1-正則化

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

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

→ 二次計画問題（スパース解）

- uLSIF:非負拘束なし, ℓ_2-正則化

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

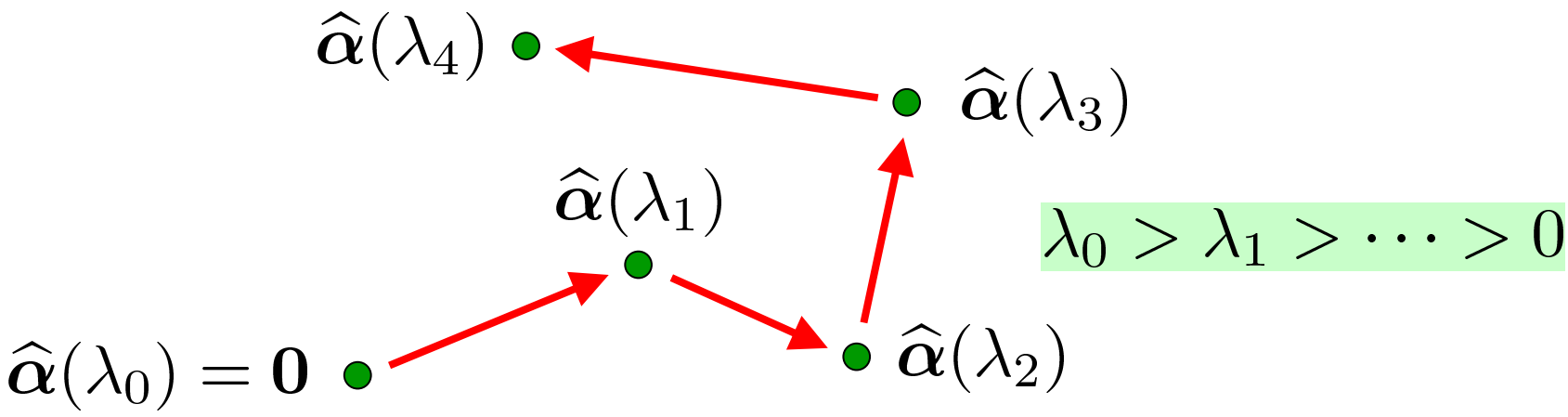
→ 解析解 $(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$

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

cLSIF: 正則化パス追跡

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

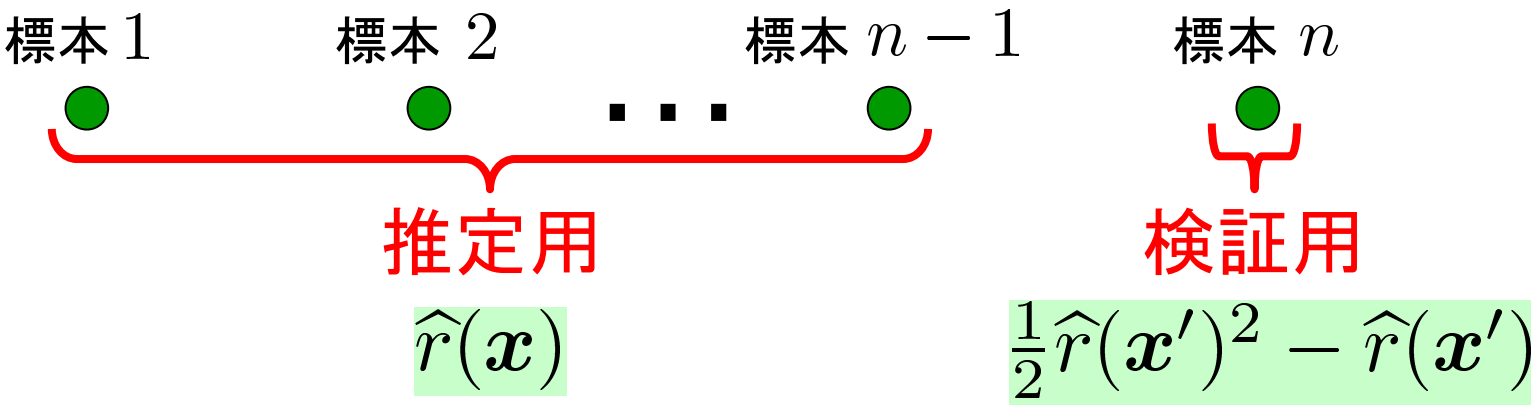
- 解は正則化パラメータ λ に関して区分線形



- 全ての λ に対する解を，二次計画問題を解くことなく効率よく計算できる

uLSIF:モデル選択

- 一つ抜き交差確認法 (LOOCV):



- LOOCVの計算は一般に n 回の繰り返しを要する
- しかし，uLSIFでは，LOOCVのスコアを解析的に計算できる（Sherman-Woodbury-Morrisonの公式を用いる）
- モデル選択の計算時間を劇的に軽減できる

発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

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

密度比適合法:統一的な枠組み

Sugiyama, Suzuki & Kanamori (IBISML2010-11)

- ブレグマン距離(BR)のもとで，密度比のモデル $r(\boldsymbol{x})$ を真の密度比 $\widehat{r}(\boldsymbol{x})$ に適合させる: $\min_{\widehat{r}} \mathrm{BR}_f(\widehat{r})$

$$\mathrm{BR}_f(\widehat{r}) = \int p_{\mathrm{de}}(\boldsymbol{x}) \nabla f(\widehat{r}(\boldsymbol{x})) \widehat{r}(\boldsymbol{x}) d\boldsymbol{x} - \int p_{\mathrm{de}}(\boldsymbol{x}) f(\widehat{r}(\boldsymbol{x})) d\boldsymbol{x}$$

$$- \int p_{\mathrm{nu}}(\boldsymbol{x}) \nabla f(\widehat{r}(\boldsymbol{x})) d\boldsymbol{x} + C$$

- BRの標本近似は次式で与えられる:

$$\mathrm{BR}_f(\widehat{r}) \approx \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \nabla f(\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}})) \widehat{r}(\boldsymbol{x}_j^{\mathrm{de}}) - \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} f(\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}}))$$

$$- \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \nabla f(\widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}})) + C$$

密度比適合法

- 様々な方法を含む統一的な密度比推定の枠組み:
  - ロジスティック回帰:

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

  - カーネル平均適合:

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

  - カルバック・ライブラー重要度推定:

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

  - 最小二乗密度比適合:

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

  - べき距離(ロバスト推定):

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

Basu, Harris, Hjort & Jones (Biometrika1998)

発表の流れ

1. 密度比推定の応用例
2. 密度比推定手法
   - A) 確率的分類法
   - B) 積率適合法
   - C) 密度適合法
   - D) 密度比適合法
   - E) 統一的な枠組み
   - F) 次元削減

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

次元削減付き密度比直接推定 (D^3; D-cube)

- 密度推定を経由せず，密度比を直接する手法は有望である.
- しかし，高次元の場合は密度比推定はやはり困難である.
- 密度比推定と次元削減を組み合わせる

異分布部分空間

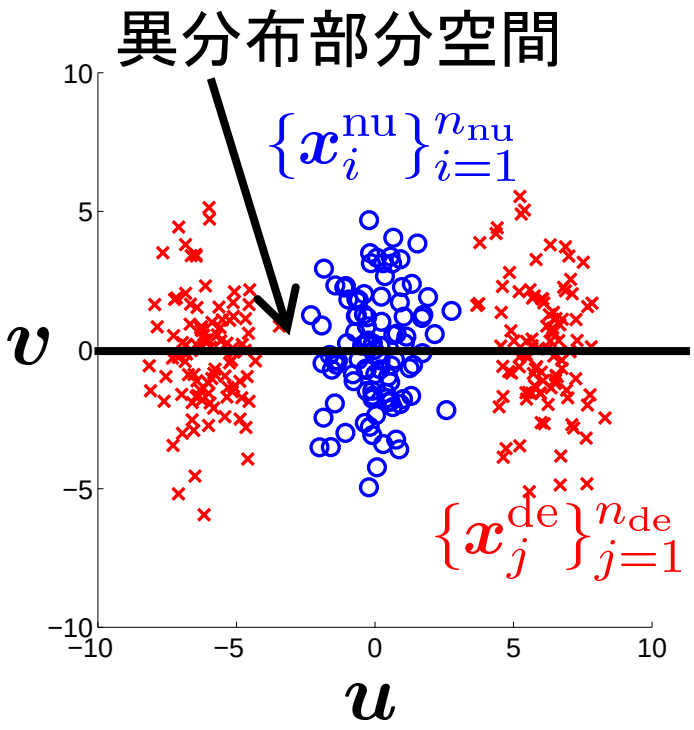
- 仮定: $p_{\mathrm{nu}}(\boldsymbol{x})$ と $p_{\mathrm{de}}(\boldsymbol{x})$ は, ある部分空間（異分布部分空間）内でのみ異なる

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

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

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

$\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$:フルランクで直交

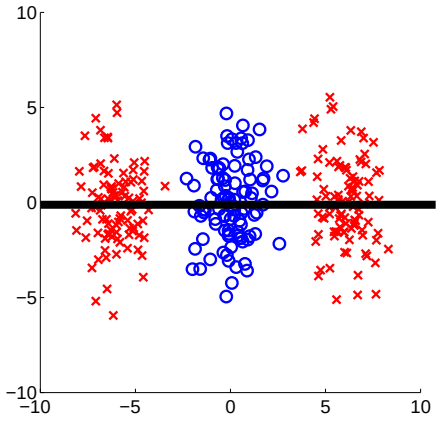


- 低次元の異分布部分空間内のみで密度比を推定できる.

教師付き次元削減による異分布部分空間の探索

Sugiyama, Kawanabe & Chui (NN2010)

- 異分布空間内では$p_{\mathrm{nu}}(\boldsymbol{x})$と$p_{\mathrm{de}}(\boldsymbol{x})$が異なるため，$\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$と$\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$は分離可能であろう.
- 任意の教師付き次元削減法によって，異分布部分空間を見つけることができるであろう.
- 局所フィッシャー判別分析(LFDA):
  - クラス内多峰性を扱うことができる.
  - 部分空間の次元数に上限がない.
  - 解析解が得られる.



Sugiyama (JMLR2007)

教師付き次元削減アプローチの利点と欠点

■ 利点:

- uLSIF+LFDAによってD^{3}の解も解析的に求められる

■ 欠点:

- $\boldsymbol{u}$ と $\boldsymbol{v}$ が独立という理論的制限がある

$$p(\boldsymbol{v}|\boldsymbol{u}) = p(\boldsymbol{v})$$

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

- 分離可能性を高めることにより異分布部分空間が見つかるとは限らない(例:重なりのある異なる分布)

分離可能 ⇒ 分布が異なる

分離可能 ⇏ 分布が異なる

異分布部分空間の特徴付け

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

- 異分布部分空間はピアソン距離(PE)の最大解:

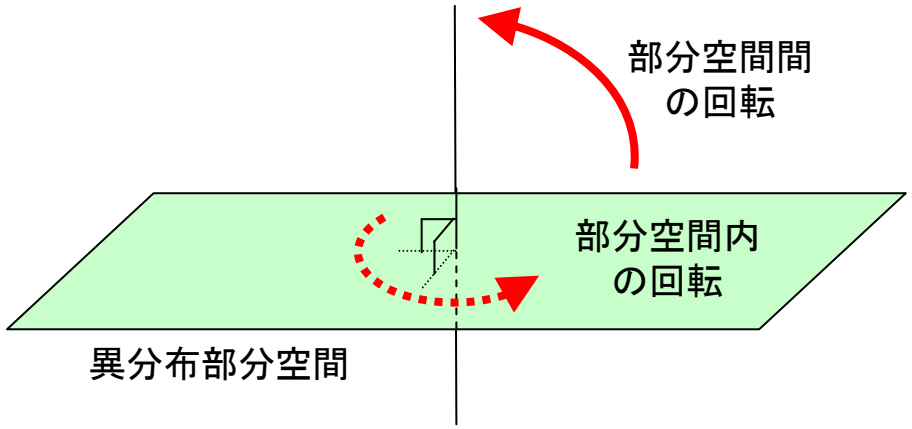
$$\mathrm{PE}[p_{\mathrm{nu}}(\boldsymbol{u}), p_{\mathrm{de}}(\boldsymbol{u})] = \int \left(\frac{p_{\mathrm{nu}}(\boldsymbol{u})}{p_{\mathrm{de}}(\boldsymbol{u})} - 1 \right)^2 p_{\mathrm{de}}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$$

- 従って, PEを $\boldsymbol{U}$ に関して最大化すれば, 異分布部分空間が求まる
- PEはuLSIFによって解析的に近似できる
  - 最小二乗異分布部分空間探索(LHSS)

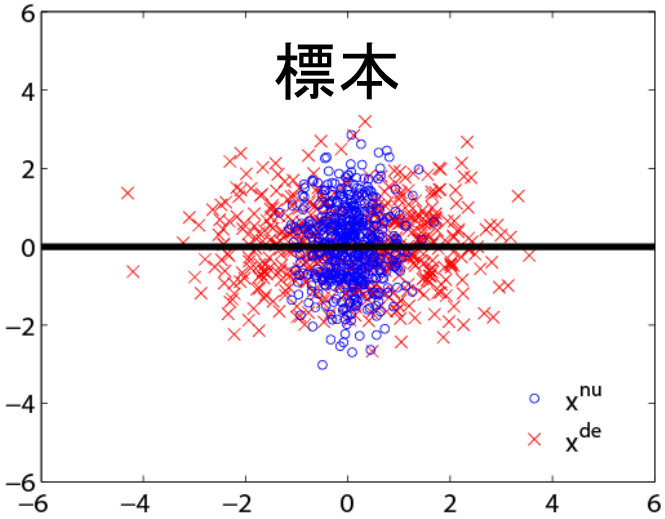
$$\widehat{\mathrm{PE}} = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \left(\widehat{r}(\boldsymbol{u}_j^{\mathrm{de}}) - 1 \right)^2 \qquad \boldsymbol{u}_j^{\mathrm{de}} = \boldsymbol{U}\boldsymbol{x}_j^{\mathrm{de}}$$

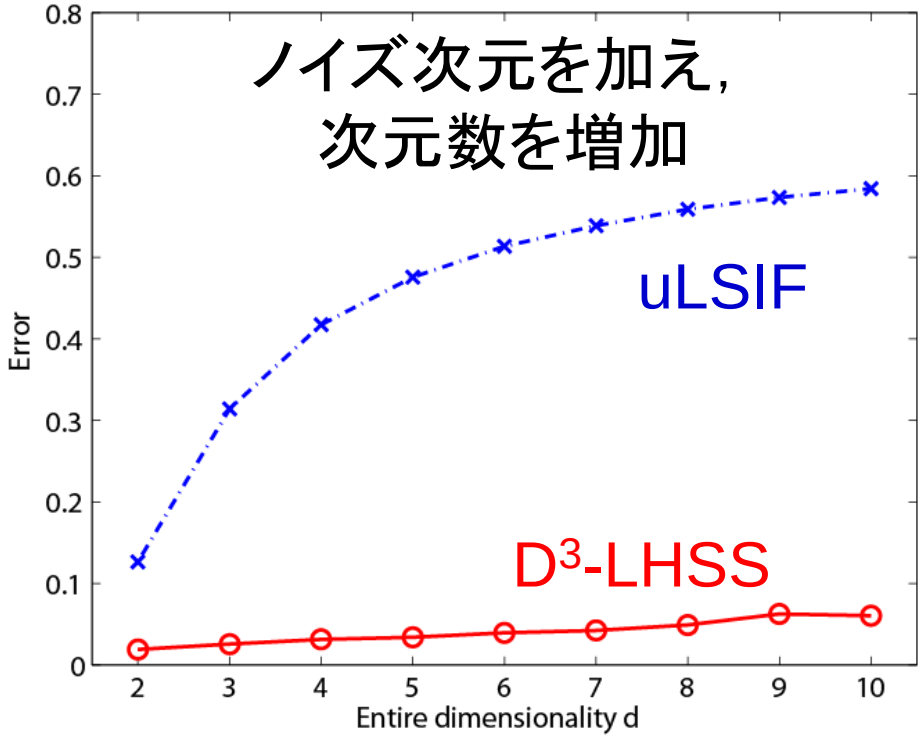
異分布部分空間の直接探索

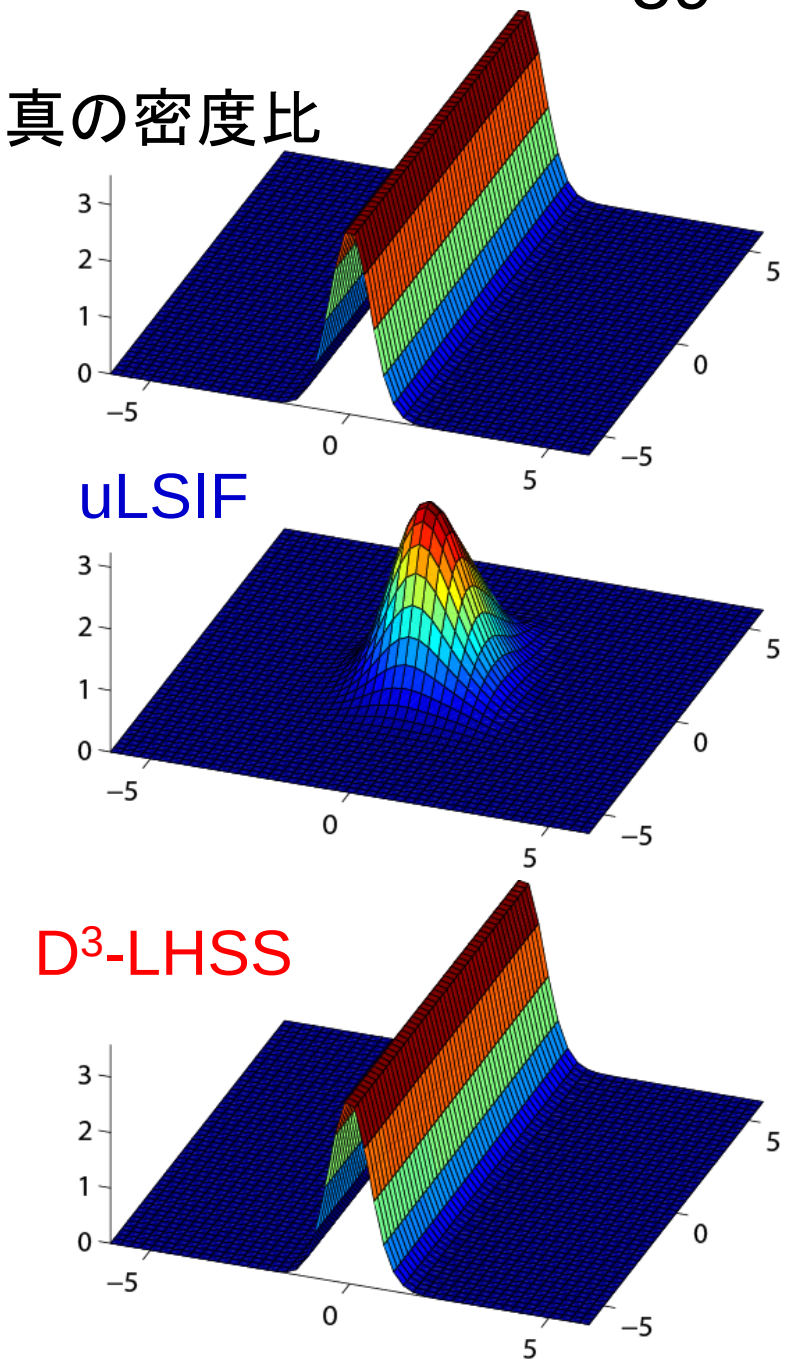
- 勾配法:
- 自然勾配法: Amari (NeuralComp1998)
  - スティーフェル多様体の構造を利用して収束性を改善
- ギブンズ回転:
  - 二つの軸を選び，その平面内で部分空間を回転
- 部分空間回転: Plumbley (NeuroComp2005)
  - 部分空間内での回転を無視することにより，計算効率を改善



数値例

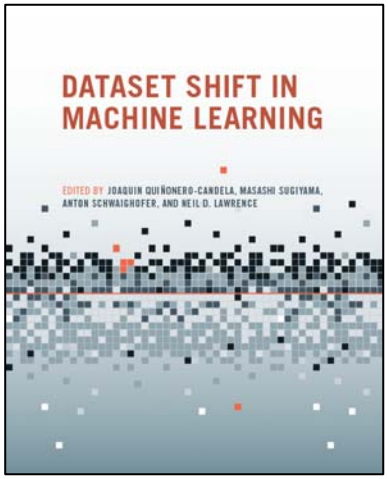

標本
x^{nu}
x^{de}


ノイズ次元を加え，
次元数を増加
uLSIF
D³-LHSS
Error
Entire dimensionality d


真の密度比
uLSIF
D³-LHSS

書籍

- Quiñonero-Candela, Sugiyama, Schwaighofer & Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009.



- Sugiyama & Kawanabe
  Covariate Shift Adaptation,
  MIT Press, 2011.

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

解説論文

- 密度比推定の枠組み:
  - 杉山，密度比に基づく機械学習の新たなアプローチ，統計数理，2010.
  - Sugiyama, Kanamori, Suzuki, Hido, Sese, Takeuchi & Wang. A density-ratio framework for statistical data processing. IPSJ Transactions on Computer Vision and Applications, vol.1, pp.183-208, 2009.
- 密度比推定法:
  - Sugiyama, Suzuki & Kanamori. Density ratio estimation: A comprehensive review. 京都大学数理解析研究所講究録, no.1703, pp.10-31, 2010.
- 共変量シフト適応:
  - 杉山，非定常環境下での教師付き学習:データの入力分布が変化する場合，画像ラボ, vol.18, no.10, pp.1-6, 2007.
  - 杉山，共変量シフト下での教師付き学習，日本神経回路学会誌, vol.13, no.3, pp.111-118, 2006.
- 論文，MATLABコードが以下からダウンロードできます:

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

まとめ:密度比の世界

実問題応用例:
ブレイン・コンピュータインターフェース，ロボット制御，音声認識，
画像認識，自然言語処理，バイオインフォマティクス，データマイニング

機械学習アルゴリズム:
重点サンプリング（共変量シフト適応，ドメイン適応，多タスク学習），
二標本問題（二標本検定，外れ値検出，変化点検知），
相互情報量推定（独立性検定，変数選択，独立成分分析，
次元削減，因果推定）
条件付き確率推定（可視化，状態遷移推定，確率的パターン認識），

密度比推定法:
基本アルゴリズム（LR，KMM，KLIEP，LSIF），
大規模対応，高次元対応，安定化，ロバスト化，統一化

理論解析:
収束性解析（確率論），情報量規準（統計学），安定性解析（最適化）