非凸学習理論チーム
Non-Convex Learning Theory Team

金森 敬文
Kanamori Takafumi

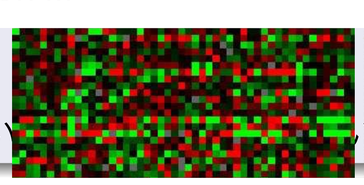
カーネル法に対する変数選択

(金森，熊谷らによる [Matsui, et al, arXiv:1806.00569]

- 多次元データを観測
- 推定・予測に影響を与える重要な変数群を選択したい
  variable selection, subset selection ...

重要な少数の要因を選択： 解釈

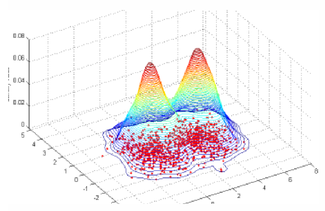
同程度の説明力 ⟹ 変数が少ない傾向.



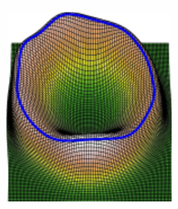
Our Goal

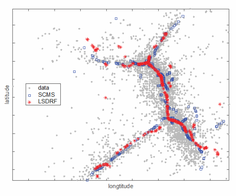
さまざまな問題設定で非線形関数の変数選択を行う .

- 線形関数の変数選択： adaptive lasso [Breiman,'95; Zou,'06; Yuan & Lin,'07].
  - 問題：(高次元) 線形回帰モデル
- 非線形関数の変数選択： 提案法 [Matsui, et al.,'18].
  - 問題： 回帰分析， 密度推定， 密度比推定，
    密度稜線推定・ モード ・ クラスタリング [Sasaki, et al.,'18],・・・

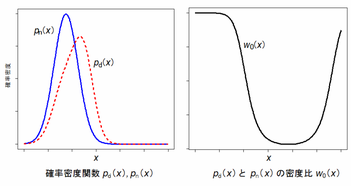


確率密度





密度稜線



密度比

- モデル： Power Series カーネル (再生核空間がスケーリングに関して不変)

理論的結果：正しい変数セットを選ぶ確率 $\longrightarrow 1$ (データ数 $\to \infty$)

例：異なる病院間のデータを (密度比でバイアス補正して) 統合.
適切な変数セットを選択できる.

転移学習の理論解析

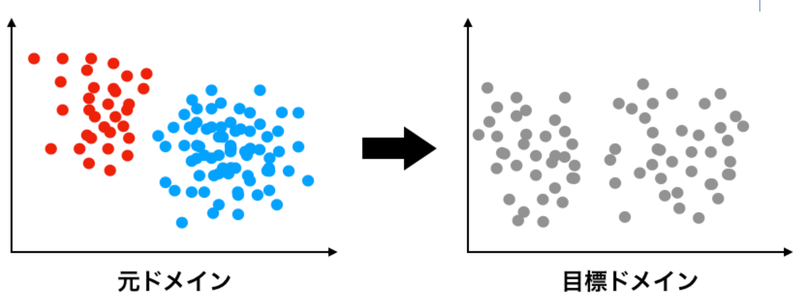
(熊谷らによる[Kumagai & Matsui, IBIS2018])

■転移学習

- 目的: 対象となるデータドメインにおいて高い性能を達成すること
- 仮定: 異なる分布を持つ他のデータドメインが利用可能
- 利点: 複数のドメインの利用によりアルゴリズムの性能が向上
- 問題: 転移学習アルゴリズムの性能の理論評価

■本研究の問題設定

- 2つのドメイン (元ドメイン，目標ドメイン) が与えられる
- サンプル空間，ラベル空間，仮説集合，損失関数は共通
- 元ドメインではラベルありデータ，目標ドメインでは
  ラベルなしデータが利用可能



相関情報を罰則に導入したスパース・モデリング

(藤澤らによる [Takada et al.])

スパース回帰における Lasso の問題点

Lassoの推定量にはバイアスが内在している. 高次元では本質的.
不要な非有効変数を取り込みやすい.

$$\beta_1^* x_1 \approx \hat{\beta}_1^{\mathrm{Lasso}} x_1 + \left(\beta_1^* - \hat{\beta}_1^{\mathrm{Lasso}}\right) x_2$$

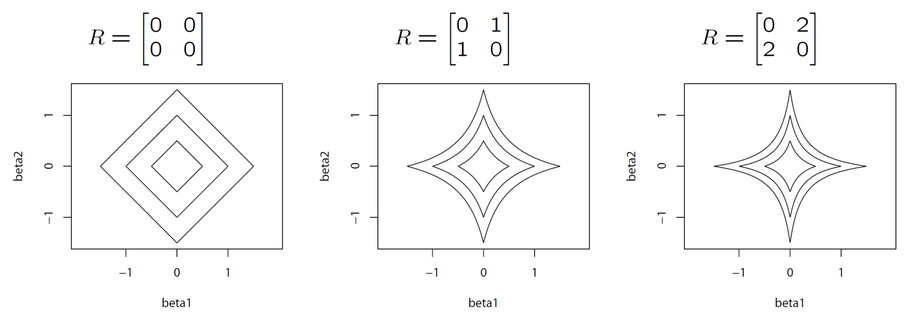
$$\hat{\beta}_1^{\mathrm{Lasso}} \neq \beta_1^* \qquad \text{when } x_1 \approx x_2$$

問題点の克服方法

相関の高い変数は同時に選ばれにくくする罰則を導入する.

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \left(\|\beta\|_1 + \frac{\alpha}{2} \sum_{j,k} R_{jk} |\beta_j||\beta_k| \right)$$

R_{jk} : 標本相関係数の絶対値の単調増加関数.
罰則第二項の効果 : R_{jk} が非常に大きいとき, β_j と β_k のどちらかを, 0にしやすい.



$\|\beta\|_1 + \frac{1}{2}|\beta|^\top R|\beta|$ $(\beta \in \mathbb{R}^2)$ の等高線

成果のポイント

パラメータ推定アルゴリズム: 座標降下法に基づいた効率的手法.
非漸近バウンドを導出: 適当な場合には提案手法はLassoより良い.
数値例で有効性を確認: SCAD・MCP・Exclusive Group Lasso より良い.
GLMにも理論結果の拡張が可能

■本研究の主結果

損失関数 ℓ はラベル空間上の距離関数であるとする. また任意の仮説 h に対し $\ell(h(x), y)$ が K-リプシッツであるとき以下が成り立つ:

$$R_T(h) \leq R_S(h) + K\mathcal{W}_P(P_S(X), P_T(X)) + K\|f_S - f_T\|_Q$$

ここで R_T, R_S は期待リスク, $\mathcal{W}_P$ は(変則的な)Wasserstein距離, $P_T(X), P_S(X)$ は周辺分布, f_T, f_S はラベル付け関数, $\|\cdot\|_Q$ はQに関するL1ノルムを表す

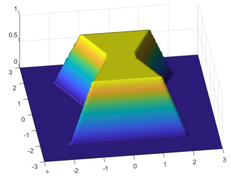
■考察

- 既存研究(Shen+ AAAI2018など)よりも評価がタイトになった
- 周辺分布間の最適輸送を学習することで上界の第2項を最小化できる
  (元ドメインのデータのみで学習したのでは最小化できない)
- 上界の第3項は目標ドメインのラベルがないため最小化できない
  → 妥当な仮定のもとでより小さくできるかは今後の課題

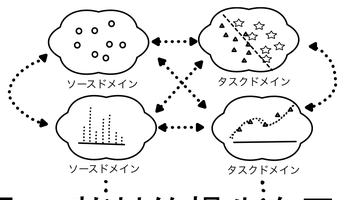
来年度に向けて

➢ 以下，投稿中・準備中

課題 1. 規格化定数の計算が困難な場合の分布推定



課題 2. 不確実性を含むマルチドメイン環境下での転移学習.



マルチドメイン，マルチソース間の双方向方学習の構築.

課題 3. 敵対的損失を用いる学習の理論を深化 (課題1と関連)

課題 4. 相互推薦システムの統計的モデリング・各種グラフノード類似度との関連