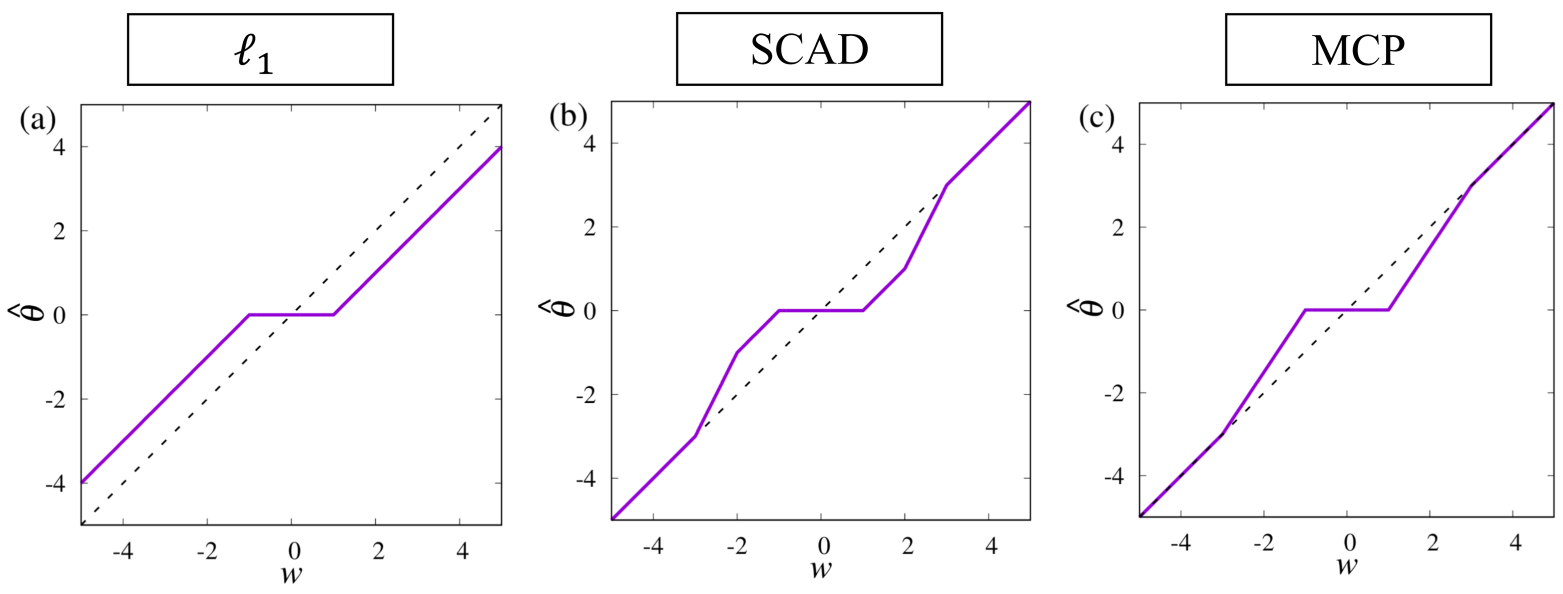
非凸スパース正則化によるモデル選択

坂田 綾香　　数理・推論研究系　助教

【導入：スパース正則化つき推定】

$$\hat{\theta} = \arg\min_{\theta}\left\{\frac{1}{2}(w-\theta)^2 + J(\theta)\right\}$$



入力データに応じて推定値がゼロ成分を取りうるような正則化Jをスパース正則化と呼ぶ。

- 変数選択(モデル選択)と推定が同時にできる。
- どういうモデルを選べば良いの？　→ 今回の研究

【問題設定】

スパース正則化つき線形回帰

$$\min_{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{y}-A\boldsymbol{x}\|_2^2 + J(\boldsymbol{x})\right\}$$

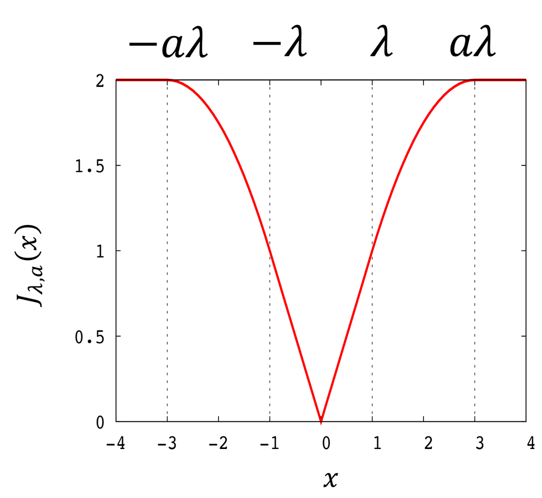
スパース正則化

- $\boldsymbol{y}\in\mathbb{R}^M$: 応答変数
- $A\in\mathbb{R}^{M\times N}$:説明変数
- $\boldsymbol{x}\in\mathbb{R}^N$:回帰係数

今回考える正則化…非凸スパース正則化

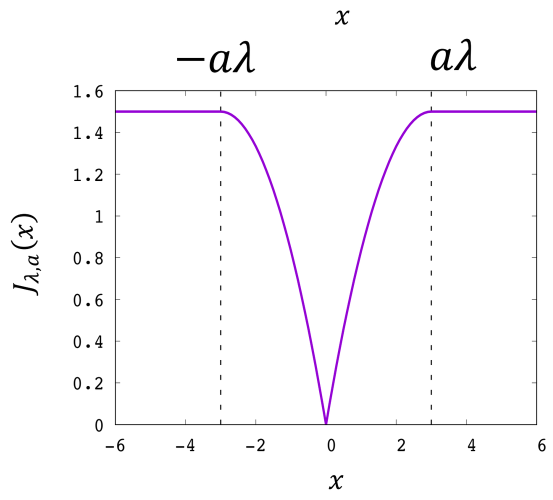
- SCAD (Smoothly Clipped Absolute Deviation)

$$J_{\lambda,a}(x) = \begin{cases} \lambda|x| & (|x|\le\lambda) \\ -\dfrac{x^2-2a\lambda|x|+\lambda^2}{2(a-1)} & (\lambda<|x|\le a\lambda) \\ \dfrac{(a+1)\lambda^2}{2} & (|x|>a\lambda) \end{cases}$$



- MCP (Minimax Concave Penalty)

$$J_{\lambda,a}(x) = \begin{cases} \lambda|x| - \dfrac{x^2}{2a} & (|x|\le a\lambda) \\ \dfrac{a\lambda^2}{2} & (|x|>a\lambda) \end{cases}$$



- これらの正則化の特徴 … 推定値が連続で不偏 (一番上の図参照)
- 正則化パラメータλ, aによって非ゼロ要素の位置・値が変わる

→ 正則化パラメータの選び方を提案します。

【方針】

予測誤差を最小化するように正則化パラメータを選ぶ

(1) In sample error: $\varepsilon_{\rm in}(\boldsymbol{y}) \equiv \dfrac{1}{N}E_{\boldsymbol{z}}[\|\boldsymbol{z}-A\hat{\boldsymbol{x}}(\boldsymbol{y})\|_2^2]$

※ $\hat{\boldsymbol{x}}(\boldsymbol{y})$: データ$\boldsymbol{y}$のもとでの推定値
$\boldsymbol{z}$: $\boldsymbol{y}$と同じ分布に従うデータ

(2) Extra sample error: $\varepsilon_{\rm ext}(\boldsymbol{y}) \equiv E_{z,\boldsymbol{a}_z}\left[\left(z-\boldsymbol{a}_z\hat{\boldsymbol{x}}(\boldsymbol{y})\right)^2\right]$

※ $\hat{\boldsymbol{x}}(\boldsymbol{y})$: データ$\boldsymbol{y}$のもとでの推定値
$\boldsymbol{a}_z$: 応答変数zに付随した説明変数

真の分布を知らなければ厳密評価できないので
推定量を構成する

(1) In sample errorについて

次の関係が成立する

$$E[\varepsilon_{\rm in}(\boldsymbol{y})] = E[\varepsilon_{\rm train}(\boldsymbol{y})] + 2\sigma_y^2{\rm df}$$

ここで

$$\varepsilon_{\rm train}(\boldsymbol{y}) = \frac{1}{M}[\|\boldsymbol{y}-A\hat{\boldsymbol{x}}(\boldsymbol{y})\|_2^2] \quad \text{… 訓練誤差(観測可)}$$

$${\rm df} = \frac{1}{M\sigma_y^2}\sum_\mu {\rm cov}[y_\mu, \hat{y}_\mu] \quad \text{… 一般化自由度(観測不可)}$$

Stein's Lemmaから推定量を構成する (データがGaussianのときは不偏)

$${\rm df} = \frac{1}{M}\sum_\mu E\left[\frac{d\hat{y}_\mu}{dy_\mu}\right] \longrightarrow \widehat{\rm df} = \frac{1}{M}\sum_\mu \frac{d\hat{y}_\mu}{dy_\mu}$$

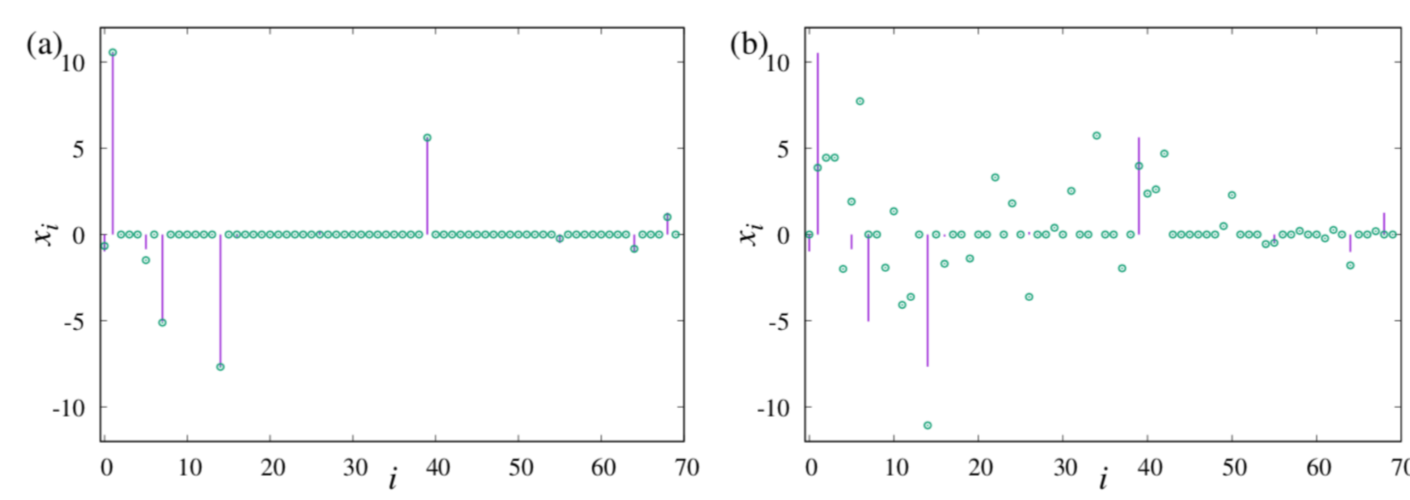
In sample errorの推定量

$$\hat{\varepsilon}_{\rm in}(\boldsymbol{y}) = \varepsilon_{\rm train}(\boldsymbol{y}) + 2\sigma_y^2\widehat{\rm df}$$

一般化自由度の推定量を、確率伝搬法を使って評価する。[参考文献1参照]

Communities and Crime Unnormalized Data Set

Our estimator　　　　AIC



- Solid line: Model with minimum true prediction error
- Circles: Models with minimum estimator

真の予測誤差最小のモデルと近いモデルを選択する。

(2) Extra sample errorについて

推定量: Leave-one-out cross validation error

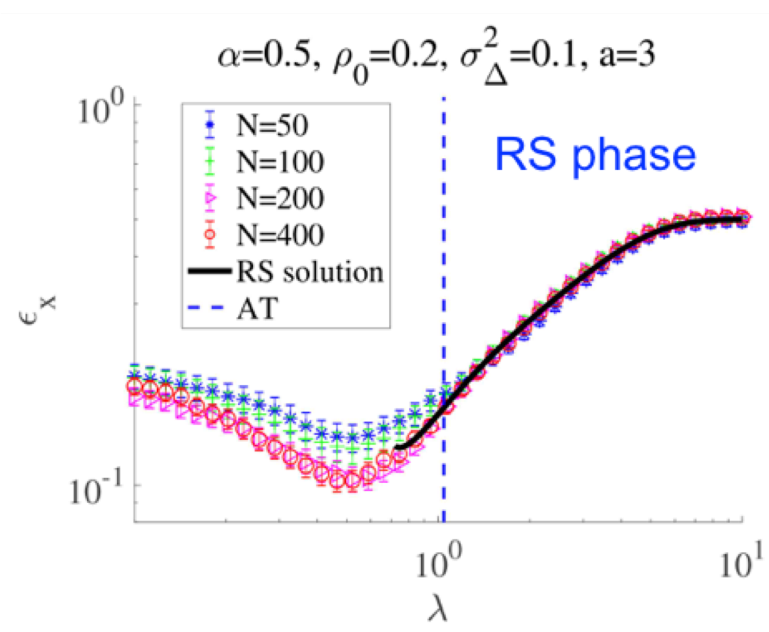
$$\hat{\varepsilon}_{\rm ext}(\boldsymbol{y}) \equiv \frac{1}{M}\sum_\mu\left(y_\mu - \boldsymbol{a}_\mu^{\rm T}\hat{\boldsymbol{x}}(\boldsymbol{y}_{\backslash\mu})\right)^2$$

$\hat{\boldsymbol{x}}(\boldsymbol{y}_{\backslash\mu})$ …μ番目のデータがない時の推定値

- 定義通りの評価にかかる計算コストが大きい。
- そこで非凸正則化にも使える近似公式を導出。
- また、モデル候補は「RS phase」と呼ばれる領域に制限すべき。

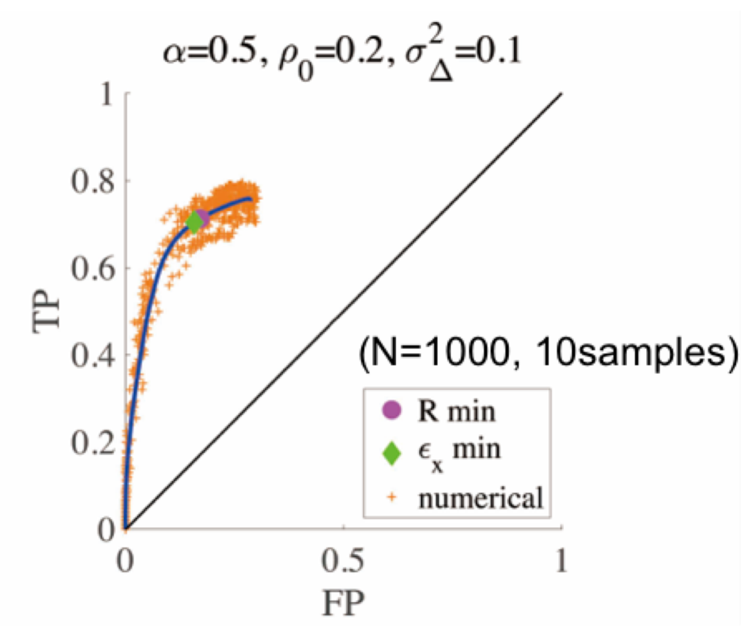
[参考文献2参照]

CV errorの理論値と数値計算結果の比較



RS領域では、
理論値と近似公式の結果が一致する。

ROC curve



ROC curve上で(0,1)と最も近い点と
CV error最小点は非常に近い。
(L1等には見られない性質)

【参考文献】

1. A. Sakata, Journal of Statistical Mechanics: Theory and Experiment, vol. 2018, 063404 (2018).
2. T. Obuchi & A. Sakata, arXiv:1902.10375.