11 Linear Regression

11.4.6 Variable selection consistency
11.4.7 Group lasso
11.4.8 Elastic net (ridge and lasso combined)

Eito/sushiyoshi

目次

Ridge回帰とLasso回帰（復習）

Ridge回帰

$$\hat{w}_{\mathrm{ridge}} = \arg\min_{w}\left\{\mathrm{RSS}(w) + \lambda\|w\|_2^2\right\}$$

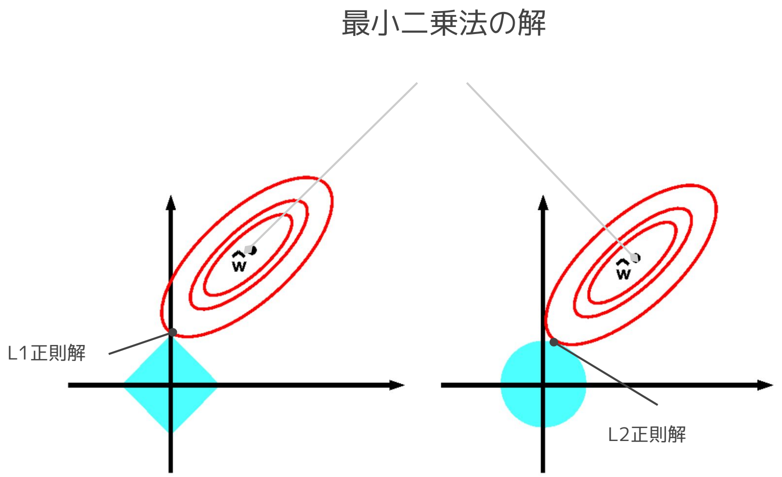
- ✓ **過学習**を防ぐ
- ✓ **L2正則化(L2ノルムの二乗和)**
- ✓ 寄与の小さい係数を**縮小**（**スパースにはならない**）

RSS(ω):残差閉包和。最小二乗法の最小化関数

Lasso回帰

$$\hat{w}_{\mathrm{lasso}} = \arg\min_{w}\left\{\mathrm{RSS}(w) + \lambda\|w\|_1\right\}.$$

- ✓ **過学習**を防ぐ
- ✓ **L1正則化(L1ノルム)**
- ✓ 寄与の小さい係数を**0にして削除**（**スパースになる**）
- ✓ **角**に最適解が存在→係数0化



L1正則化

$\|w\|_1 = |w_1| + |w_2| \le B$

L2正則化

$\|w\|_2^2 = w_1^2 + w_2^2 \le B$

形状の違いがスパース性を生む

Debiasing（復習）

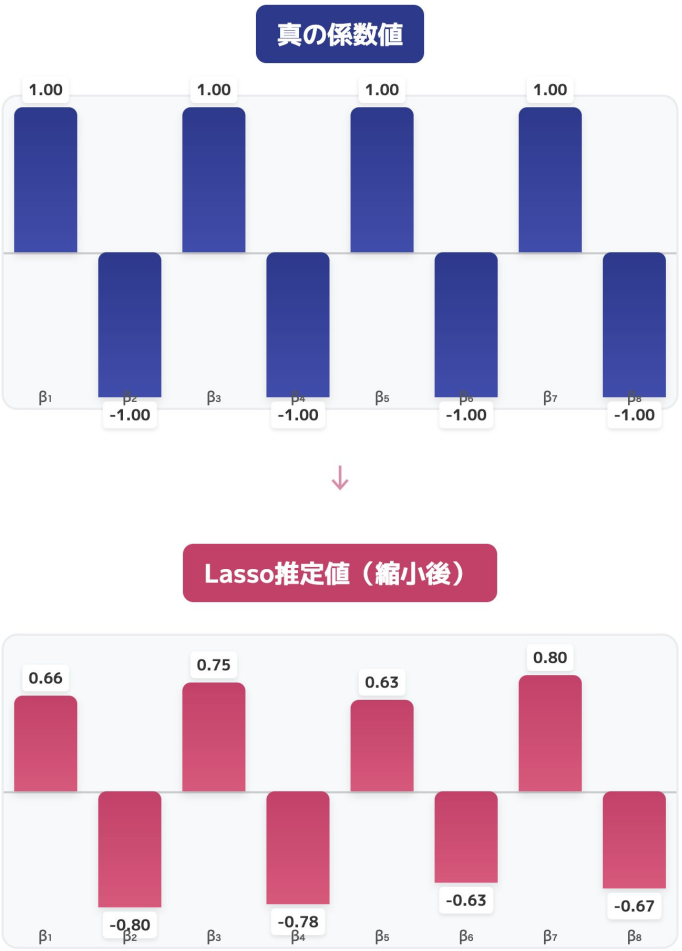
Lassoの問題点

Lassoは重要な係数も0に向けて**縮小**してしまう
→ 真の係数値よりも**小さく推定される**

Lasso係数のDebiasing

以下の2段階の推定を行う。
1. **Lasso回帰**で重要な変数を推定
2. 選択された変数だけで**最小二乗法**
(Ordinary Least Squares:OLS)

OLSの**過学習**と、
Lassoの**係数縮小**を相互にカバー



目次

- 復習
    - Ridge回帰とLasso回帰
    - Debiasing
- 11.4.6 Variable selection consistency
    - 変数選択の一貫性
    - Example
- 11.4.7 Group lasso
    - Group lasso
    - Group lassoの応用例
    - Group Lassoの正則化項(L2ノルム)
    - Group lassoの正則化項(無限大ノルム)
    - L2ノルムのGroup sparsity
    - 無限大ノルムのGroup sparsity
    - Example
- 11.4.8 Elastic net (ridge and lasso combined)
    - Elastic net
    - Elastic netの利点

変数選択の一貫性

変数選択(variable selection)

L1正則化によって**重要な説明変数**のみを抽出

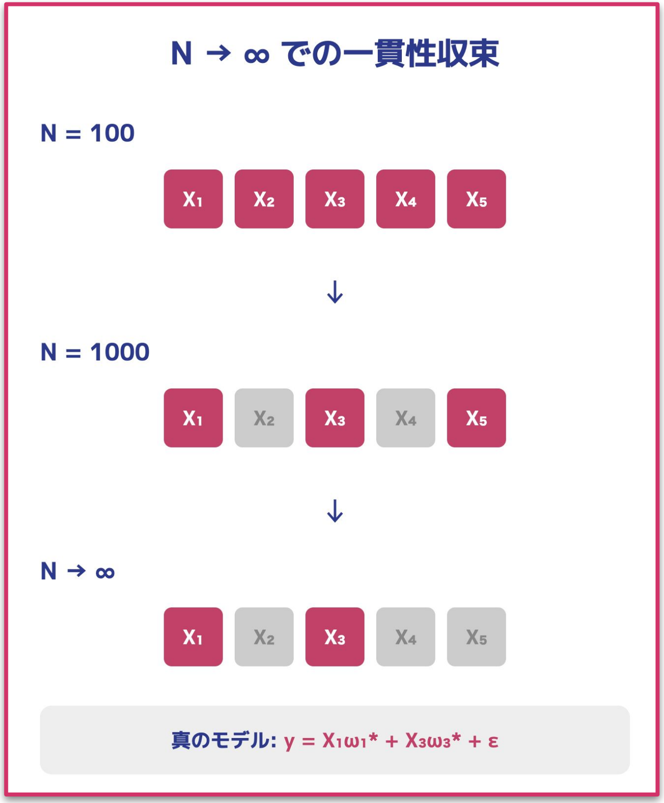
✔ **解釈性** ↑ ✔ **汎化性能** ↑ ✔ **計算コスト** ↓

変数選択の一貫性(model selection consistent)

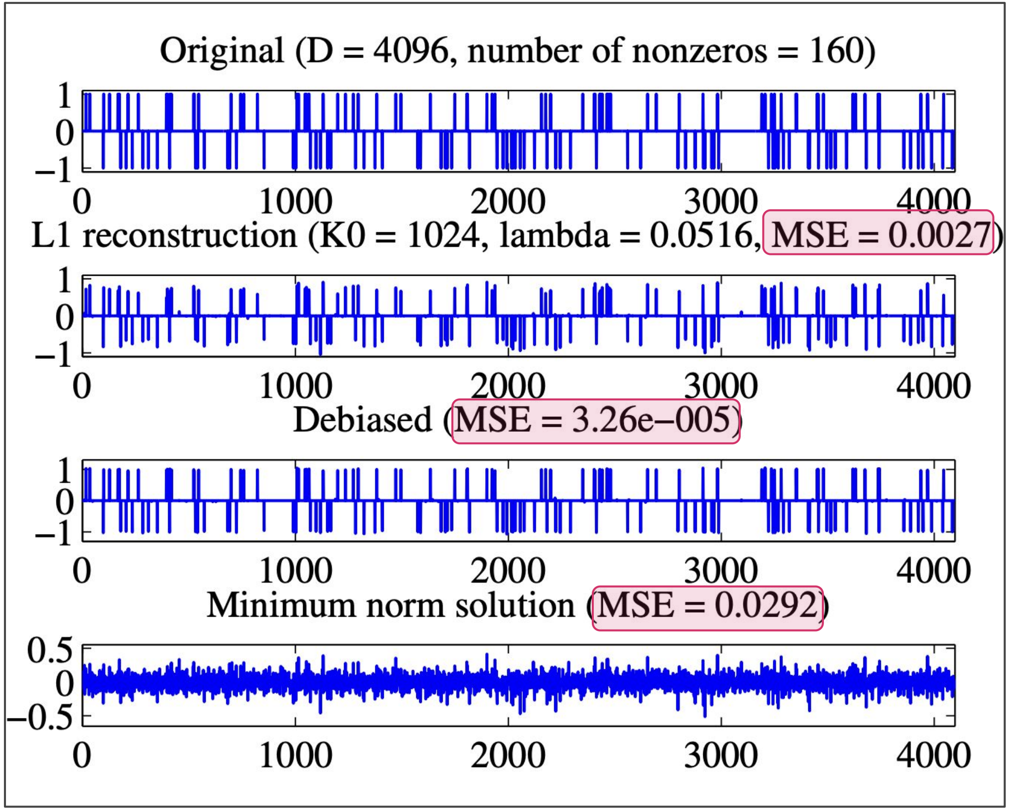
$N \to \infty$で推定された $\hat{\omega}$が**真のスパース解** ω^*に一致する性質

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\omega}^* + \boldsymbol{\epsilon}$$

$\boldsymbol{y}$	：目的変数ベクトル	N	：データのサンプル数
$\boldsymbol{X}$	：$N \times D$ の説明変数行列	D	：説明変数の数
$\boldsymbol{\omega}^*$	：真のパラメータベクトル		
$\boldsymbol{\epsilon}$	：ω^*による予測誤差で、$\varepsilon_n \sim \mathcal{N}(0, 0.01^2)$		



Example



説明変数の数 D=4096、
データのサンプル数 N=1024
± 1 スパイクをランダムに160個配置

1. Original
2. Lasso回帰
3. Debiased(Lasso→OLS)
4. OLS

MSE(平均二乗誤差)は、
Debiased < Lasso回帰 < OLS

目次

- 復習
  - Ridge回帰とLasso回帰
  - Debiasing
- 11.4.6 Variable selection consistency
  - 変数選択の一貫性
  - Example
- 11.4.7 Group lasso
  - Group lasso
  - Group lassoの応用例
  - Group Lassoの正則化項(L2ノルム)
  - Group lassoの正則化項(無限大ノルム)
  - L2ノルムのGroup sparsity
  - 無限大ノルムのGroup sparsity
  - Example
- 11.4.8 Elastic net (ridge and lasso combined)
  - Elastic net
  - Elastic netの利点

Group Lasso

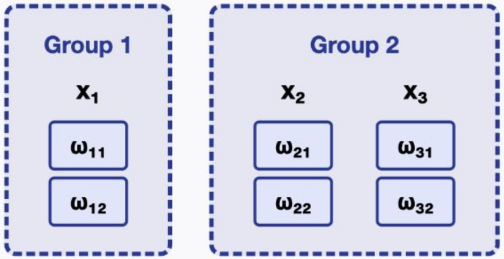
従来のLasso

- ✓ 各変数 xiとパラメータ ωiが**1対1対応**
- ✓ 変数 xiを**一個ずつ**選択して除外

正則化項: $\lambda\ \Sigma|\omega_i|$

Group Lasso

- ✓ 各変数 xiにパラメータベクトル$\boldsymbol{\omega_i = [\omega_{i,1}, \ldots, \omega_{i,D}]}$
- ✓ 変数を**グループ単位**で選択して除外（グループスパース性）



正則化項: $\lambda\ \Sigma||\omega_g||_2$

Group Lassoの応用例

カテゴリカル変数を持つ線形回帰

カテゴリカル変数はone-hotベクトル(値が0か1のみの離散的なベクトル)で表現されるため、変数を除外する際はベクトル全体を0にする。

例：都道府県（47次元）を除外する場合、47個すべての重みを同時に0にする

多項ロジスティック回帰

各変数がクラス数分の重みを持つため、変数を除外する際は全クラスの重みベクトルを0にしする。

例：3クラス分類で特徴量を除外する場合、3つの重みを同時に0にする

ニューラルネットワーク

ある特定のニューロンを「オフ」にするには、そのニューロンへの入力重みをすべて0にする。

例：隠れ層のニューロンを無効化し、ネットワーク構造を単純化

マルチタスク学習

各特徴量がタスク数分の重みを持つため、すべてのタスクで使う/使わないを一括で決定できる。

例：複数の予測タスクで共通して重要な特徴量を選択

Group Lassoの正則化項(L2ノルム)

$$
\begin{aligned}
\mathrm{PNLL}(\boldsymbol{w}) &= \mathrm{NLL}(\boldsymbol{w}) + \lambda \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_2 \\
&= \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \boxed{\lambda \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_2} + \left(\frac{N}{2}\log(2\pi\sigma^2)\right)
\end{aligned}
$$

ただし、

$$
\|\boldsymbol{w}_g\|_2 = \sqrt{\sum_{d\in g} w_d^2}
$$

以下は、L2ノルムがグループ単位の除外を引き起こす仕組みである。

- ✓ L2ノルムは、各グループのベクトルが成す**球の半径をなるべく小さくする**ことを要求する正則化
- ✓ 半径を小さくするために、グループの各成分は**同じ割合で縮小される**
- ✓ グループに0要素がある場合、**他の要素も0へ**。グループベクトルは**原点へと引き寄せられる。**

また、以下の点に注意する。

- ✓ リッジ回帰のL2正則化は**L2ノルムの二乗和**であり、区別する
- ✓ **グループごとにルートで区切っている**点が大きな差異

Ridge回帰のL2正則化(比較用)

$$
\begin{aligned}
\mathrm{PNLL}(\boldsymbol{w}) &= \mathrm{NLL}(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2 \\
&= \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2 + \left(\frac{N}{2}\log(2\pi\sigma^2)\right)
\end{aligned}
$$

L2ノルムのGroup sparsity

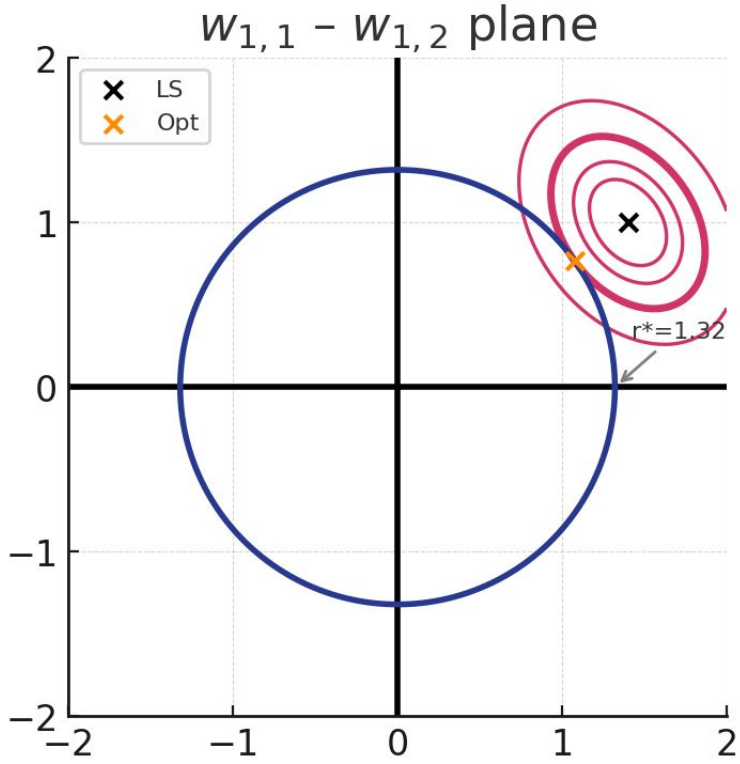
$$\frac{\partial}{\partial w_1}\left(w_1^2 + w_2^2\right)^{\frac{1}{2}} = \frac{w_1}{\sqrt{w_1^2 + w_2^2}}$$

勾配による「連帯責任」

- 勾配は半径で正規化した方向ベクトル。
- どちらか一方が 0 付近だと
  ∂/∂w ≈ ±1 → λ 倍の押し戻しで**両方 0**へ。
- 両方が大きいと勾配 ≈ 0 → ペナルティ弱く**両方残る。**

➡ グループのパラメータ全体が生死を共有。

L2ノルムのGroup sparsity(Example)



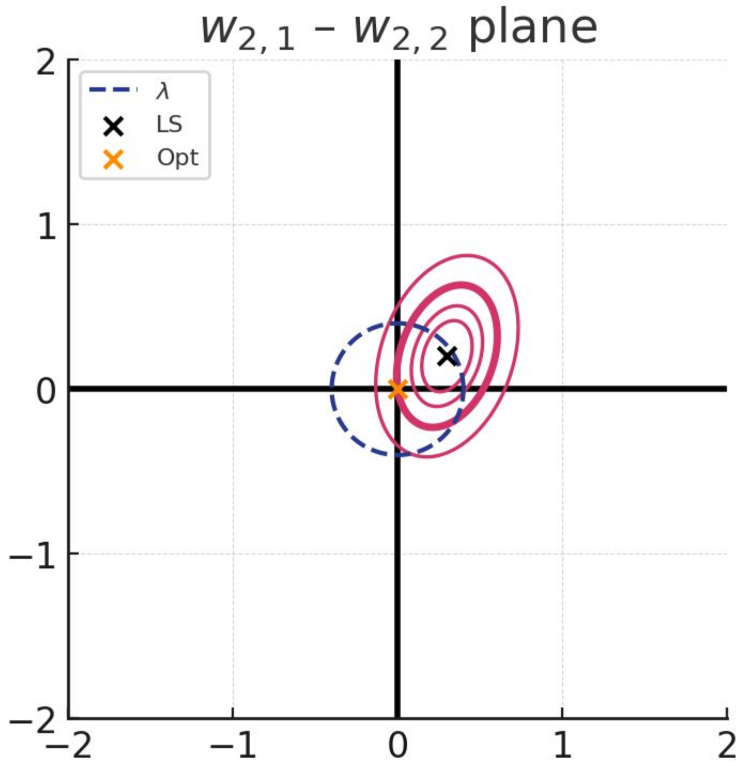
単純化のために、次のような極端な例を考える。
4つのパラメータ $w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2}$ を、
次の2つのグループに分ける。

$$\boldsymbol{w_1} = [w_{1,1}, w_{1,2}], \boldsymbol{w_2} = [w_{2,1}, w_{2,2}]$$

グループ①：円境界で縮む

- ✓ 左図は $\omega_{1,1} - \omega_{1,2}$ 平面である。
- ✓ ピンクは誤差関数の等高線
- ✓ 濃紺の円はL2正則化項の制約条件
- ✓ LS 解（黒×）は円の外 → ペナルティで一律縮小
- ✓ 最適解（橙×）は円周と等高線の接点

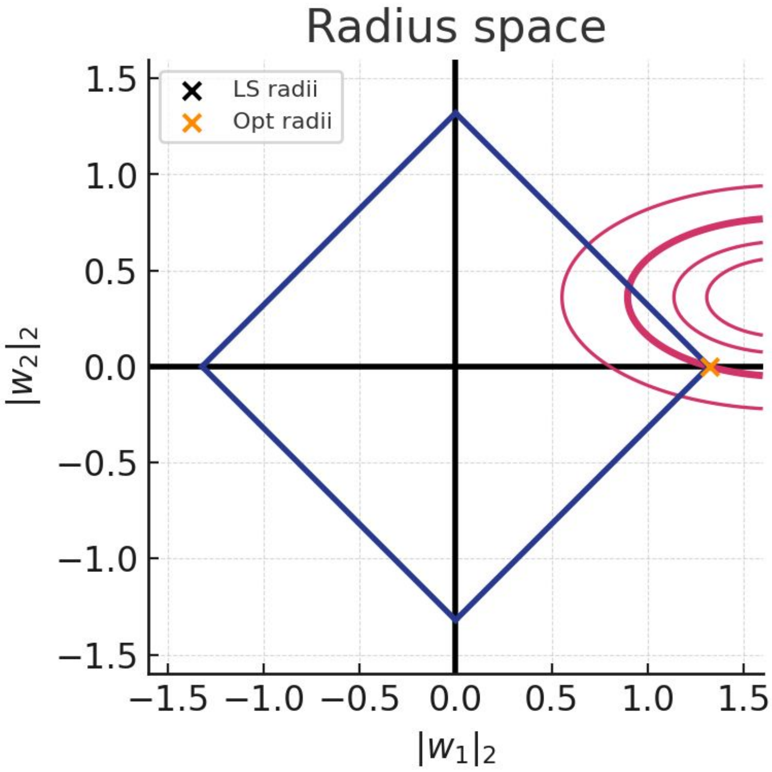
L2ノルムのGroup sparsity(Example)



グループ②：閾値以下なら 0

- 左図は $\omega_{2,1}$ ー $\omega_{2,2}$平面である。
- 破線円は閾値 \(\lambda=0.4\)
- LS 解は円内 → **blocksoft-threshold** で原点へ
- 橙× = 黒× = (0,0) → グループごと消滅

L2ノルムのGroup sparsity(Example)



半径空間で見るブロック選択

- 左図は $\|\boldsymbol{w_1}\|_2 - \|\boldsymbol{w_2}\|_2$ 平面
- 菱形： $\|\boldsymbol{w_1}\|_2 + \|\boldsymbol{w_2}\|_2 = t$
- 接点は右端角 ⇒ $\|\boldsymbol{w_2}\|_2 = 0$
- グループ1のみ採択、グループ2は 0

Group Lassoの正則化項(無限大ノルム)

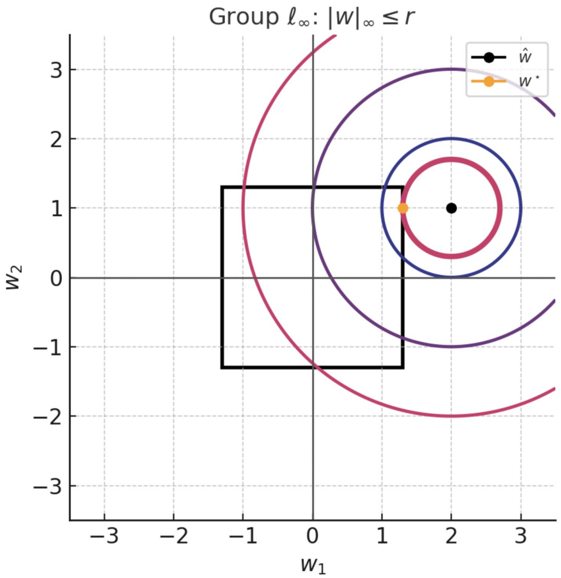
$$\mathrm{PNLL}(\boldsymbol{w}) = \mathrm{NLL}(\boldsymbol{w}) + \lambda \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_\infty$$

$$= \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \boxed{\lambda \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_\infty} + \left(\frac{N}{2}\log(2\pi\sigma^2)\right)$$

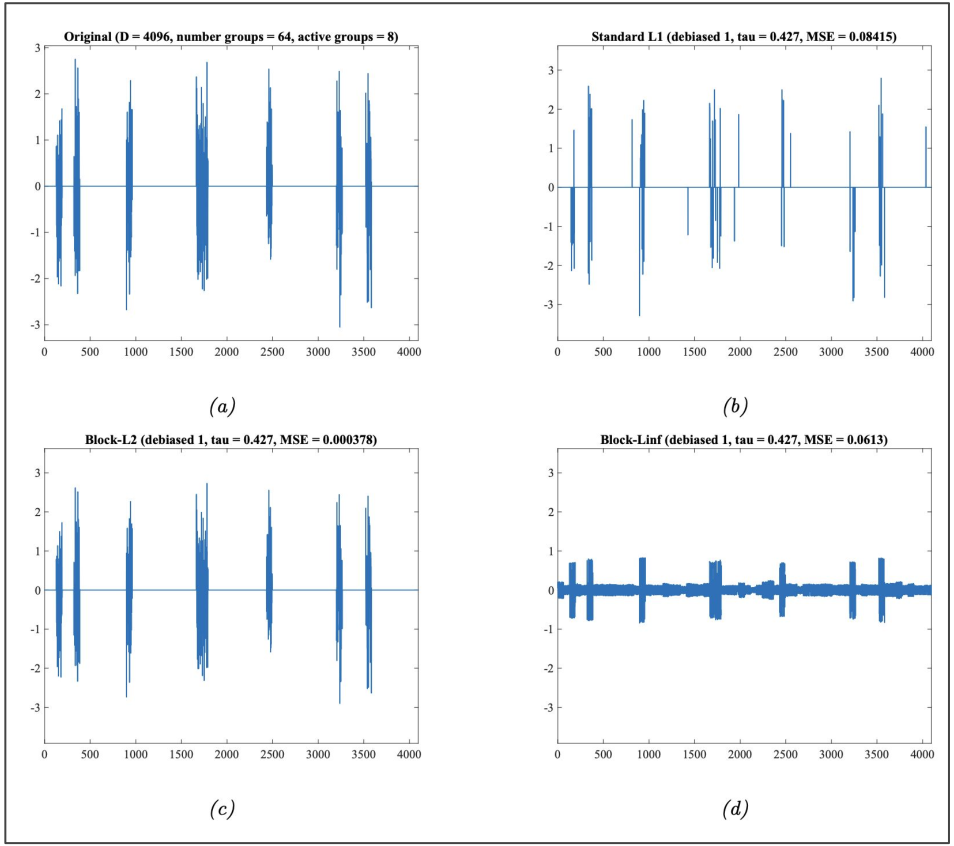
ただし、

$$\|\mathbf{w}_g\|_\infty = \max_{d \in g} |w_d|$$

- ✓ グループごとに最大値をとる
- ✓ グループの中で重要な特徴が1つでもあれば残す



Example



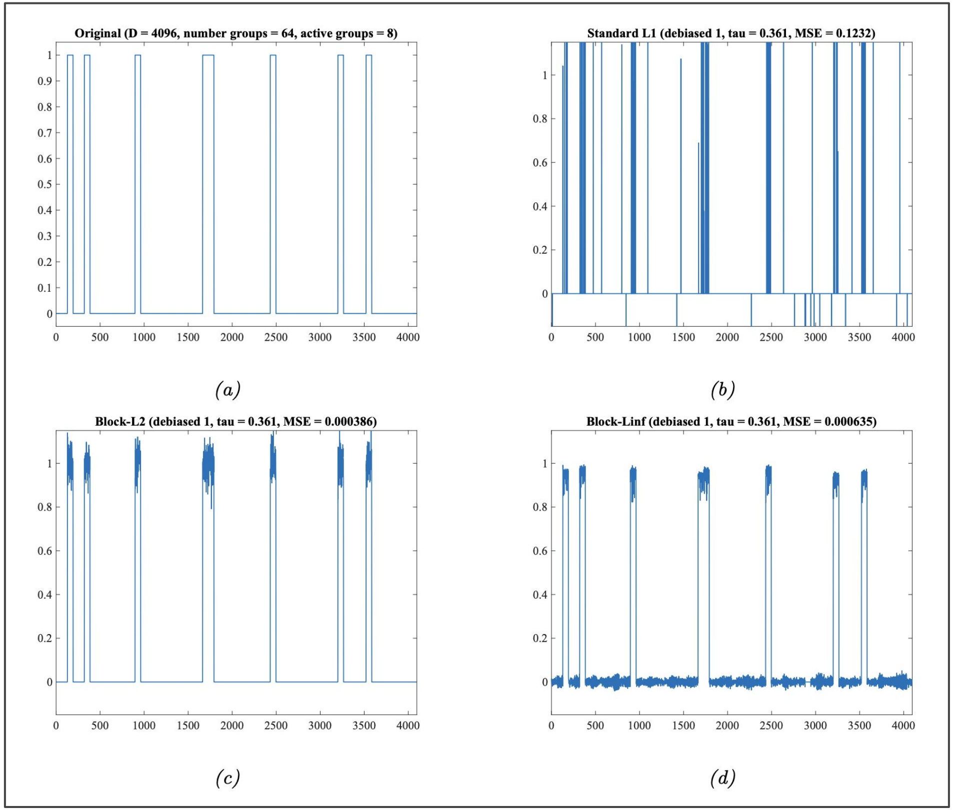
説明変数の数 D=4096、
データのサンプル数 N=1024
64グループに分割して8グループ選択
$\omega \in \mathcal{N}(0,1);$
いずれもLasso→OLSでDebiasを行う

a. Original
b. Lasso
c. Group Lasso(2-norm)
d. Group Lasso(infinity-norm)

MSE(平均二乗誤差)は、
2-norm < infinity-norm < Lasso

Debiasedはグループ構造を無視。
2-normは元信号を**ほぼ完全に回復**。
infinity-normは過度に平坦化。

Example



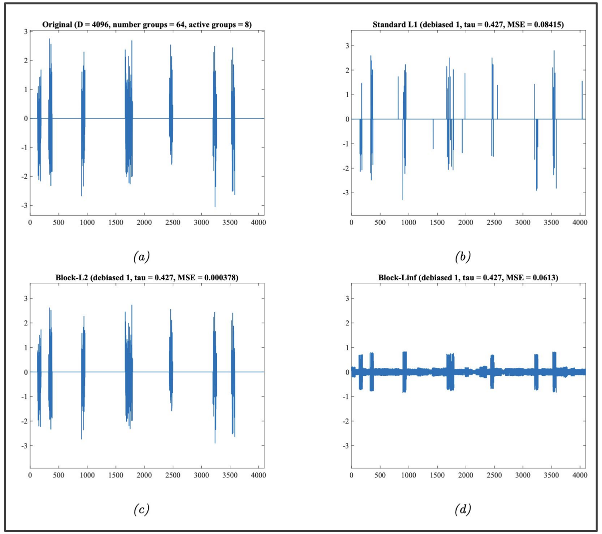
説明変数の数 D=4096、
データのサンプル数 N=1024
64グループに分割して8グループ選択
$\omega \in \{0,1\}$
いずれもLasso→OLSでDebiasを行う。

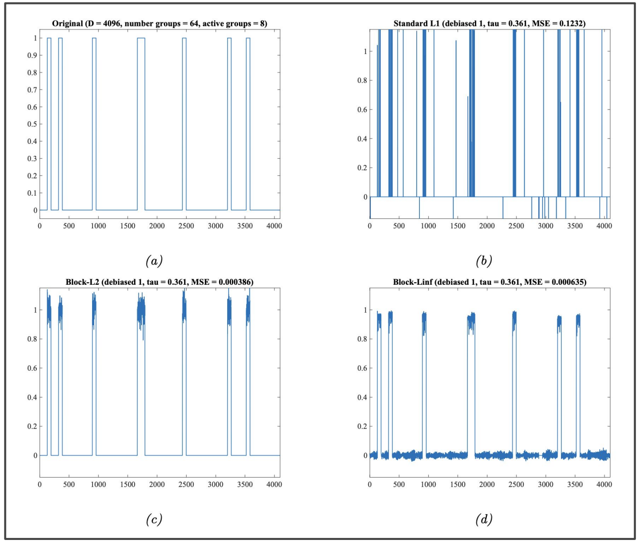
a. Original
b. Lasso
c. Group Lasso(L2-norm)
d. Group Lasso(infinity-norm)

MSE(平均二乗誤差)は、
2-norm < infinity-norm < Debiased
変わらず。

しかし、infinity-normは
「**ブロック内を同じ大きさにする**」
特性により
L2-normよりも綺麗に回復

Example





$$\omega \in \mathcal{N}(0,1);$$

「**元の値を綺麗に回復する**」
特性により、L2-normが最適。

$$\omega \in \{0,1\}$$

「**ブロック内を同じ大きさにする**」
特性により、infinity-normが最適。

目次

- 復習
  - Ridge回帰とLasso回帰
  - Debiasing
- 11.4.6 Variable selection consistency
  - 変数選択の一貫性
  - Example
- 11.4.7 Group lasso
  - Group lasso
  - Group lassoの応用例
  - Group Lassoの正則化項(L2ノルム)
  - Group lassoの正則化項(無限大ノルム)
  - L2ノルムのGroup sparsity
  - 無限大ノルムのGroup sparsity
  - Example
- 11.4.8 Elastic net (ridge and lasso combined)
  - Elastic net
  - Elastic netの利点

Elastic net

なぜ Elastic Net か？

- ✓ **Ridge** は**多重共線性**に強いが、*変数選択*はできない。
- ✓ **Lasso** は**スパース性**を得られるが、*相関の高い特徴*を同時に残しにくい。
- ✓ **Elastic Net** はRidgeのL2正則化とLassoのL1正則化の利点を総取りした手法

Group Lasso との対比

- ✓ Group Lasso は **グループ構造を事前指定 する**必要あり。
- ✓ 未知でも「**相関の高い係数が自然にグループ化**」されてほしい。
- ✓ そこで Ridge 成分が相関係数を束ね、Lasso 成分が不要なグループを消去 → **Elastic Net** が最適。

Elastic net

$$L(w, \lambda_1, \lambda_2) = \|y - Xw\|^2 + \lambda_2 \|w\|_2^2 + \lambda_1 \|w\|_1$$

記号	意味
$X \in \mathbb{R}^{N \times D}$	説明変数行列（N サンプル × D 特徴）
$y \in \mathbb{R}^{N}$	目的変数ベクトル
$w \in \mathbb{R}^{D}$	回帰係数（学習対象）
$\|\cdot\|_2^2$	L_2 ノルム（二乗）— Ridge 部分
$\|\cdot\|_1$	L_1 ノルム — Lasso 部分
λ_1	Lasso 強度 (スパース性 ↑)
λ_2	Ridge 強度 (共線性対策 ↑)

項ごとの役割

- ▶ 残差二乗和 … 当てはまりの良さ
- ▶ L_1 項 … 係数そのものを 0 へ → 変数選択
- ▶ L_2 項 … 係数を 滑らか に → 相関特徴のグループ保持