スパース推定の数理：統計理論から計算手法まで

†‡ 鈴木　大慈

† 東京工業大学
‡ JST さきがけ

2015 年 12 月 24 日
応用数理学会「科学技術計算と数値解析」研究部会セミナー

Outline

1. スパース推定のモデル
2. いろいろなスパース正則化
3. スパース推定の理論
   - $n \gg p$ の理論
   - $n \ll p$ の理論
   - より高度なトピック
4. 高次元線形回帰の検定
5. スパース推定の最適化手法

高次元データでの問題意識

- ゲノムデータ
- 金融データ
- 協調フィルタリング
- コンピュータビジョン
- 音声認識

次元 $d = 10000$ の時，サンプル数 $n = 1000$ で推定ができるか？
どのような条件があれば推定が可能か？

何らかの**低次元性 (スパース性)** を利用.

歴史: スパース推定の手法と理論

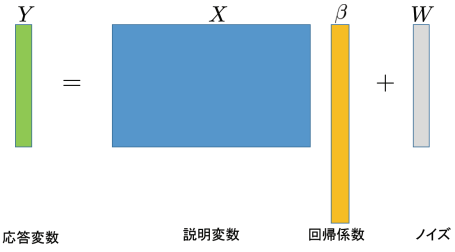
1992	**Donoho and Johnstone**	**Wavelet shrinkage (Soft-thresholding)**
1996	**Tibshirani**	**Lasso の提案**
2000	**Knight and Fu**	**Lasso の漸近分布 ($n \gg p$)**
2006	**Candes and Tao, Donoho**	**圧縮センシング (制限等長性，完全復元，$p \gg n$)**
2009	**Bickel et al., Zhang**	**制限固有値条件 (Lasso のリスク評価，$p \gg n$)**
2013	**van de Geer et al., Lockhart et al.**	**スパース推定における検定 ($p \gg n$)**

これ以前にも反射法地震探査や画像雑音除去，忘却付き構造学習に $L1$ 正則化は使われていた．詳しくは田中利幸 (2010) を参照.

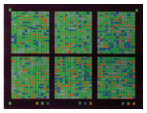
Outline

1. スパース推定のモデル
2. いろいろなスパース正則化
3. スパース推定の理論
   - $n \gg p$ の理論
   - $n \ll p$ の理論
   - より高度なトピック
4. 高次元線形回帰の検定
5. スパース推定の最適化手法

高次元データ解析



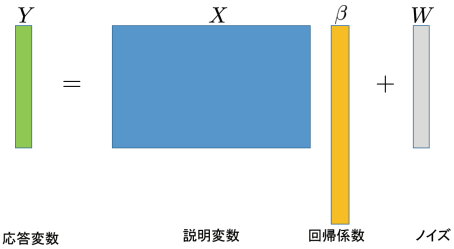
サンプル数 $\ll$ 次元


バイオインフォ

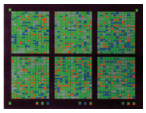

テキストデータ


画像データ

高次元データ解析



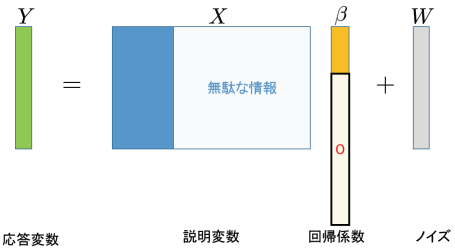
サンプル数 $\ll$ 次元
$\times$ 古典的数理統計学：サンプル数 $\gg$ 次元


バイオインフォ


テキストデータ


画像データ

スパース推定



サンプル数 $\ll$ 次元
無駄な情報を切り落とす→スパース性

Lasso 推定量

R. Tsibshirani (1996). Regression shrinkage and selection via the lasso.
J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267–288.
引用数：14728 (2015 年 12 月 23 日)

変数選択の問題（線形回帰）

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n\times p}$, 応答変数 $Y = (Y_i) \in \mathbb{R}^n$
p (次元) $\gg n$ (サンプル数).
真のベクトル $\beta^* \in \mathbb{R}^p$: 非ゼロ要素の個数がたかだか d 個 (スパース).

$$\text{モデル：}\quad Y = X\beta^* + \epsilon$$
$$(Y_i = \textstyle\sum_{j=1}^p X_{ij}\beta_j^* + \epsilon_i) \quad (i = 1, \dots, n))$$

(Y, X) から β^* を推定.
実質推定しなくてはいけない変数の数は d 個→変数選択.

変数選択の問題（線形回帰）

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n\times p}$, 応答変数 $Y = (Y_i) \in \mathbb{R}^n$
p (次元) $\gg n$ (サンプル数).
真のベクトル $\beta^* \in \mathbb{R}^p$: 非ゼロ要素の個数がたかだか d 個 (スパース).

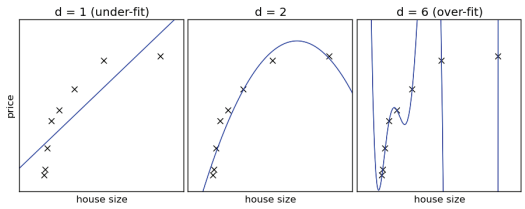
$$\text{モデル：}\quad Y = X\beta^* + \epsilon$$
$$(Y_i = \textstyle\sum_{j=1}^p X_{ij}\beta_j^* + \epsilon_i) \quad (i = 1, \dots, n))$$

(Y, X) から β^* を推定.
実質推定しなくてはいけない変数の数は d 個→変数選択.

Mallows' C_p, AIC:

$$\hat\beta_{\mathrm{MC}} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 + 2\sigma^2\|\beta\|_0$$

ただし $\|\beta\|_0 = |\{j \mid \beta_j \neq 0\}|$.
→ 2^p 個の候補を探索. NP-困難.



http://www.astroml.org/sklearn_tutorial/practical.html

$$y = b + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d + \epsilon$$

Lasso 推定量

Mallows' C_p 最小化: $\hat\beta_{\mathrm{MC}} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 + 2\sigma^2\|\beta\|_0$.

問題点: $\|\beta\|_0$ は凸関数ではない．連続でもない．沢山の局所最適解.
→ 凸関数で近似.

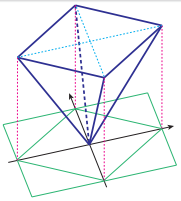
Lasso [L_1 正則化]

$$\hat\beta_{\mathrm{Lasso}} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 + \lambda\|\beta\|_1$$

ただし $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$.

→ 凸最適化！

- L_1 ノルムは L_0 ノルムの $[-1, 1]^p$ における凸包 (下から抑える最大の凸関数)
- L_1 ノルムは要素数関数の **Lovász** 拡張



Lasso 推定量のスパース性

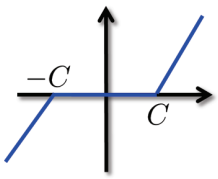
$p = n$, $X = I$ の場合.

$$\hat{\beta}_{\mathrm{Lasso}} = \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \quad \frac{1}{2}\|Y - \beta\|^2 + C\|\beta\|_1$$

$$\Rightarrow \quad \hat{\beta}_{\mathrm{Lasso},i} = \underset{b \in \mathbb{R}}{\mathrm{argmin}} \quad \frac{1}{2}(y_i - b)^2 + C|b|$$

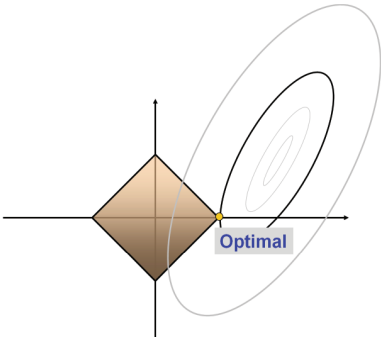
$$= \begin{cases} \mathrm{sign}(y_i)(y_i - C) & (|y_i| > C) \\ 0 & (|y_i| \leq C). \end{cases}$$

小さいシグナルは **0** に縮小される→スパース！



Lasso 推定量のスパース性

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$



スパース性の恩恵

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

Theorem (Lasso の収束レート)

ある条件のもと，定数 C が存在して高い確率で次の不等式が成り立つ：

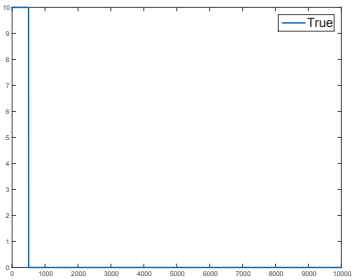
$$\|\hat{\beta} - \beta^*\|_2^2 \leq C\frac{d\log(p)}{n}.$$

※次元が高くても，たかだか $\log(p)$ でしか効いてこない．実質的な次元 d が支配的.
（「ある条件」については後で詳細を説明）

数値例

$$Y = X\beta + \epsilon.$$

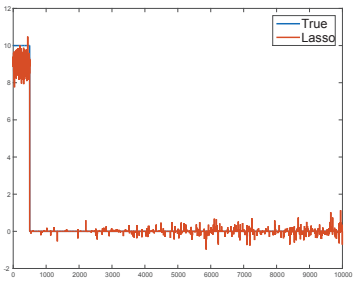
$n = 1,000$, $p = 10,000$, $d = 500$.



数値例

$$Y = X\beta + \epsilon.$$

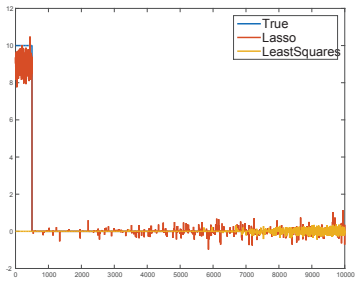
$n = 1,000$, $p = 10,000$, $d = 500$.



数値例

$$Y = X\beta + \epsilon.$$

$n = 1,000$, $p = 10,000$, $d = 500$.



Outline

1. スパース推定のモデル
2. いろいろなスパース正則化
3. スパース推定の理論
   - $n \gg p$ の理論
   - $n \ll p$ の理論
   - より高度なトピック
4. 高次元線形回帰の検定
5. スパース推定の最適化手法

Lasso を一般化

Lasso:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \underbrace{\|\beta\|_1}_{\text{正則化項}}.$$

Lasso を一般化

Lasso:

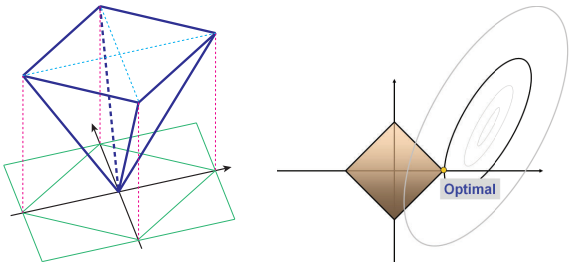
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \underbrace{\|\beta\|_1}_{\text{正則化項}}.$$

一般化したスパース正則化推定法:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\ell(z_i, \beta) + \psi(\beta).$$

$L1$ 正則化項以外にどのような正則化項が有用であろうか？

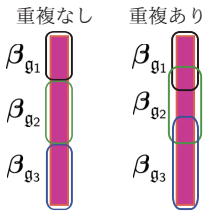
L_1 正則化によってスパースになる理由：
座標軸に沿って 尖っている.



正則化項の 尖り方 を工夫することで様々なスパース性が得られる.

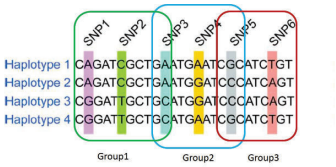
グループ正則化

$$C\sum_{\mathfrak{g}\in\mathfrak{G}}\|\beta_{\mathfrak{g}}\|$$



- グループ内すべての変数が同時に 0 になりやすい.
- より積極的にスパースにできる.

応用例：ゲノムワイド相関解析



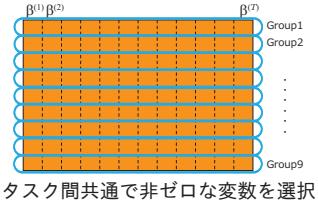
グループ正則化の応用例

- マルチタスク学習 Lounici et al. (2009)

T 個のタスクで同時に推定:

$$y_i^{(t)} \approx x_i^{(t)\top}\beta^{(t)} \quad (i = 1, \ldots, n^{(t)},\ t = 1, \ldots, T).$$

$$\min_{\beta^{(t)}} \sum_{t=1}^{T}\sum_{i=1}^{n^{(t)}}(y_i - x_i^{(t)\top}\beta^{(t)})^2 + C\underbrace{\sum_{k=1}^{p}\|(\beta_k^{(1)}, \ldots, \beta_k^{(T)})\|}_{\text{グループ正則化}}.$$



タスク間共通で非ゼロな変数を選択

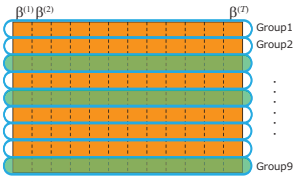
グループ正則化の応用例

- マルチタスク学習 Lounici et al. (2009)

T 個のタスクで同時に推定:

$$y_i^{(t)} \approx x_i^{(t)\top}\beta^{(t)} \quad (i=1,\dots,n^{(t)},\ t=1,\dots,T).$$

$$\min_{\beta^{(t)}} \sum_{t=1}^{T}\sum_{i=1}^{n^{(t)}}(y_i - x_i^{(t)\top}\beta^{(t)})^2 + C\underbrace{\sum_{k=1}^{P}\|(\beta_k^{(1)},\dots,\beta_k^{(T)})\|}_{\text{グループ正則化}}.$$



タスク間共通で非ゼロな変数を選択

トレースノルム正則化

$W : M \times N$ 行列.

$$\|W\|_{\mathrm{Tr}} = \mathrm{Tr}[(WW^\top)^{\frac{1}{2}}] = \sum_{j=1}^{\min\{M,N\}} \sigma_j(W)$$

$\sigma_j(W)$ は W の j 番目の特異値 (非負とする).

- 特異値の和 = 特異値への L_1 正則化 → 特異値がスパース
- 特異値がスパース = 低ランク

例: 推薦システム

	映画 A	映画 B	映画 C	⋯	映画 X
ユーザ 1	4	8	*	⋯	2
ユーザ 2	2	*	2	⋯	*
ユーザ 3	2	4	*	⋯	*
⋮					

(e.g., Srebro et al. (2005), NetFlix Bennett and Lanning (2007))

例: 推薦システム

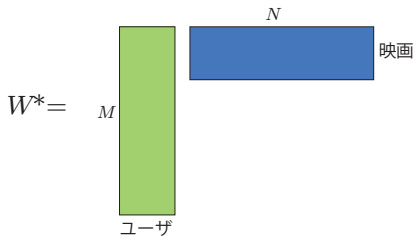
ランク 1 と仮定する

	映画 A	映画 B	映画 C	⋯	映画 X
ユーザ 1	4	8	4	⋯	2
ユーザ 2	2	4	2	⋯	1
ユーザ 3	2	4	2	⋯	1
⋮					

(e.g., Srebro et al. (2005), NetFlix Bennett and Lanning (2007))

$$\sum_{(i,j)\in T}(Y_{i,j} - W_{i,j})^2 + \lambda\|W\|_{\mathrm{Tr}}$$

例: 推薦システム



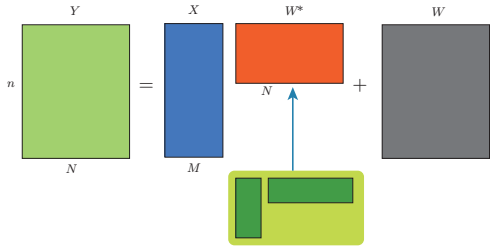
→ 低ランク行列補完:

- 低ランク行列の Rademacher Complexity: Srebro et al. (2005).
- Compressed sensing: Candès and Tao (2009), Candès and Recht (2009).

例: 縮小ランク回帰

- 縮小ランク回帰 (Anderson, 1951, Burket, 1964, Izenman, 1975)
- マルチタスク学習 (Argyriou et al., 2008)

縮小ランク回帰



W^* は 低ランク.

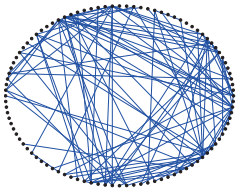
スパース共分散選択

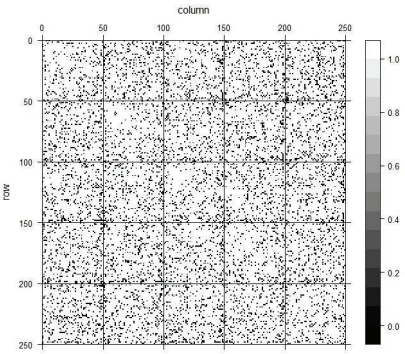
$x_k \sim N(0, \Sigma)$ (i.i.d., $\Sigma \in \mathbb{R}^{p\times p}$), $\quad \widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} x_k x_k^\top$.

$$\hat{S} = \underset{S:\text{半正定対称}}{\operatorname{argmin}} \left\{ -\log(\det(S)) + \operatorname{Tr}[S\widehat{\Sigma}] + \lambda \sum_{i,j=1}^{p} |S_{i,j}| \right\}.$$

(Meinshausen and B uhlmann, 2006, Yuan and Lin, 2007, Banerjee et al., 2008)

- Σ の逆行列 S を推定.
- $S_{i,j} = 0 \Leftrightarrow X_{(i)}, X_{(j)}$ が条件付き独立.
- ガウシアングラフィカルモデルが凸最適化で推定できる.



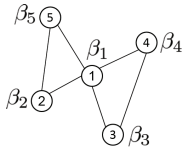


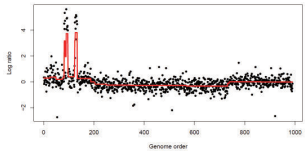
NASDAQ 銘柄からランダム抽出した 50 銘柄.
株価データを用いた分散共分散選択. 時間差も考慮.
(2011 年 1 月 4 日から 2014 年 12 月 31 日まで)
(Lie Michael, Bachelor thesis)

(一般化) Fused Lasso

$$\psi(\beta) = C \sum_{(i,j)\in E} |\beta_i - \beta_j|.$$

(Tibshirani et al. (2005), Jacob et al. (2009))

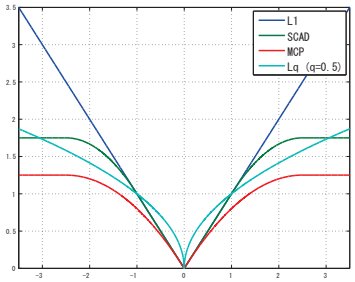




Fused lasso による遺伝子データ解析 (Tibshirani and Taylor '11)

TV デノイジング (Chambolle '04)

非凸正則化



- SCAD (Smoothly Clipped Absolute Deviation) (Fan and Li, 2001)
- MCP (Minimax Concave Penalty) (Zhang, 2010)
- L_q 正則化 ($q < 1$), Bridge 正則化 (Frank and Friedman, 1993)

よりスパースな解. その代わり最適化は難しくなる.

その他 L_1 正則化の拡張

- **Adaptive Lasso** (Zou, 2006)
  ある一致推定量 $\tilde{\beta}$ があるとして, それを利用.

$$\psi(\beta) = C \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}$$

  - Lasso よりも小さいバイアス (漸近不偏).
  - オラクルプロパティ.

- スパース加法モデル (Hastie and Tibshirani, 1999, Ravikumar et al., 2009)
  $f(x) = \sum_{j=1}^{p} f_j(x_j)$ なる非線形関数を推定.
  $f_j \in \mathcal{H}_j$ ($\mathcal{H}_j$: 再生核ヒルベルト空間) とする.

$$\psi(f) = C \sum_{j=1}^{p} \|f_j\|_{\mathcal{H}_j}$$

  - Group Lasso の一般化.
  - Multiple Kernel Learning とも呼ばれる.

Outline

1. スパース推定のモデル
2. いろいろなスパース正則化
3. スパース推定の理論
   - $n \gg p$ の理論
   - $n \ll p$ の理論
   - より高度なトピック
4. 高次元線形回帰の検定
5. スパース推定の最適化手法

問題設定

簡単のため線形回帰を考える．

$$Y = X\beta^* + \epsilon.$$

$Y \in \mathbb{R}^n$: 応答変数, $X \in \mathbb{R}^{n\times p}$: 説明変数, $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^\top \in \mathbb{R}^n$.

一般化線形回帰への一般化も可能．

$n \gg p$ の理論

Lasso の漸近分布

p は固定，$n \to \infty$ の漸近的振る舞いを考える．

- $\frac{1}{n}X^\top X \xrightarrow{\mathrm{P}} C \succ O$.
- ノイズ ϵ_i は平均 0 分散 σ^2 とする．

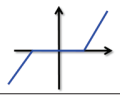
Theorem (Lasso の漸近分布 (Knight and Fu, 2000))

$\lambda_n\sqrt{n} \to \lambda_0 \geq 0$ なら

$$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{\mathrm{d}} \operatorname*{argmin}_u V(u),$$

ただし，$V(u) = u^\top Cu - 2u^\top W + \lambda_0 \sum_{j=1}^p [u_j \mathrm{sign}(\beta_j^*)\mathbf{1}(\beta_j^* \neq 0) + |u_j|\mathbf{1}(\beta_j^* = 0)]$, $W \sim N(\mathbf{0}, \sigma^2 C)$.

- $\hat{\beta}$ は $\sqrt{n}$-consistent である．
- $\beta_j^* = 0$ なる成分で $\hat{\beta}_j$ は正の確率で 0 となる．
- 第三項のせいで，漸近的にバイアスが残る．



$\hat{\beta} = \mathrm{argmin}_\beta \frac{1}{n}\|Y - X\beta\|^2 + \lambda_n \sum_{j=1}^p |\beta_j|$.

Adaptive Lasso のオラクルプロパティ

$\tilde{\beta}$ はある一致推定量．

$$\hat{\beta} = \operatorname*{argmin}_\beta \frac{1}{n}\|Y - X\beta\|^2 + \lambda_n \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}.$$

Theorem (Adaptive Lasso のオラクルプロパティ (Zou, 2006))

$\lambda_n\sqrt{n} \to 0$, $\lambda_n n^{\frac{1+\gamma}{2}} \to \infty$ のとき，

1. $\lim_{n\to\infty} P(\hat{J} = J) \to 1 \quad (\hat{J} := \{j \mid |\hat{\beta}_j| \neq 0\}, J := \{j \mid |\beta_j^*| \neq 0\})$,
2. $\sqrt{n}(\hat{\beta}_J - \beta_J^*) \xrightarrow{\mathrm{d}} N(0, \sigma^2 C_{JJ}^{-1})$.

- 変数選択の一致性あり．
- 漸近不偏，漸近正規性あり．
- ただし，β^* のある成分が原点に近づくような局所的な議論 ($\beta_j^* = O(1/\sqrt{n})$ なる状況) については何も言っていないことに注意．

$n \ll p$ の理論

Lass のリスクの上界

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^p |\beta_j|.$$

Theorem (Lasso の収束レート (Bickel et al., 2009, Zhang, 2009))

デザイン行列が *Restricted eigenvalue condition (Bickel et al., 2009)* かつ $\max_{i,j}|X_{ij}| \leq 1$ を満たし，ノイズが $\mathrm{E}[e^{\tau\epsilon_i}] \leq e^{\sigma^2\tau^2/2}$ $(\forall\tau > 0)$ を満たすなら，確率 $1-\delta$ で

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C\frac{d\log(p/\delta)}{n}.$$

- 次元が高くても，たかだか $\log(p)$ でしか効いてこない．実質的な次元 d が支配的．
- ノイズの条件はサブガウシアンの必要十分条件．

Lasso の minimax 最適性

Theorem (ミニマクス最適レート (Raskutti and Wainwright, 2011))

ある条件のもと，確率 1/2 以上で，

$$\min_{\hat{\beta}:\text{推定量}} \max_{\beta^*:d\text{-スパース}} \|\hat{\beta}-\beta^*\|^2 \geq C\frac{d\log(p/d)}{n}.$$

Lasso は minimax レートを達成する ($\frac{d\log(d)}{n}$ の項を除いて).

この結果を Multiple Kernel Learning に拡張した結果: Raskutti et al. (2012), Suzuki and Sugiyama (2013).

制限固有値条件 (Restricted eigenvalue condition)

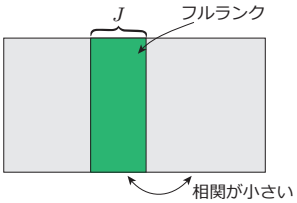
$A = \frac{1}{n}X^\top X$ とする.

Definition (制限固有値条件 ($\mathrm{RE}(k', C)$))

$$\phi_{\mathrm{RE}}(k', C) = \phi_{\mathrm{RE}}(k', C, A) := \inf_{\substack{J\subseteq\{1,\dots,n\}, v\in\mathbb{R}^p: \\ |J|\leq k', C\|v_J\|_1\geq\|v_{J^c}\|_1}} \frac{v^\top Av}{\|v_J\|_2^2}$$

に対し，$\phi_{\mathrm{RE}} > 0$ が成り立つ.

ほぼスパースなベクトルに制限して定義した最小固有値.



適合性条件 (Compatibility condition)

$A = \frac{1}{n}X^\top X$ とする.

Definition (適合性条件 ($\mathrm{COM}(J, C)$))

$$\phi_{\mathrm{COM}}(J, C) = \phi_{\mathrm{COM}}(J, C, A) := \inf_{\substack{v\in\mathbb{R}^p: \\ C\|v_J\|_1\geq\|v_{J^c}\|_1}} k\frac{v^\top Av}{\|v_J\|_1^2}$$

に対し，$\phi_{\mathrm{COM}} > 0$ が成り立つ.

$|J| \leq k'$ なら，RE よりも弱い条件.

制限等長性条件 (Restricted isometory condition)

Definition (制限等長性条件 ($\mathrm{RI}(k', \delta)$) (Candes and Tao, 2005))

ある $1 > \delta > 0$ に対し，

$$(1-\delta)\|\beta\|_2 \leq \|X\beta\|_2 \leq (1+\delta)\|\beta\|_2$$

が全ての k'-スパースなベクトル $\beta \in \mathbb{R}^p$ に対して成り立つ.

- Johnson-Lindenstrauss の補題.
- 圧縮センシングにおける完全復元の文脈でよく用いられる.

各条件の関係と収束レート

$\hat{\beta}$: Lasso 推定量.
$J := \{j \mid \beta_j^* \neq 0\}$. $d := |J|$.

		$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2$	$\|\hat{\beta}-\beta^*\|_2^2$	$\|\hat{\beta}-\beta^*\|_1^2$
強い		圧縮センシングにおける完全復元		
$\mathrm{RI}(Cd, \delta)$ + 制限直交性	→			
⇓				
$\mathrm{RE}(2d, 3)$	→	$\frac{d\log(p)}{n}$	$\frac{d\log(p)}{n}$	$\frac{d^2\log(p)}{n}$
⇓				
$\mathrm{COM}(J, 3)$	→	$\frac{d\log(p)}{n}$	$\frac{d^2\log(p)}{n}$	$\frac{d^2\log(p)}{n}$
弱い				

関連事項の詳細は Bühlmann and van de Geer (2011) に網羅されている.
完全復元は制限直交性 (Candes and Tao, 2005) なしでも示せる (Candès, 2008).
RI と RE の関係は Rudelson and Zhou (2013) でも論じられている.

制限固有値条件 (RE) が成立する確率

制限固有値条件はどれだけ成り立ちやすいか？

- p 次元確率変数 Z が等方的: $\mathrm{E}[\langle Z, z\rangle^2] = \|z\|_2^2$ $(\forall z \in \mathbb{R}^p)$.
- サブガウシアンノルム $\|Z\|_{\psi_2}$ を次のように定義する:

$$\|Z\|_{\psi_2} = \sup_{z\in\mathbb{R}^p, \|z\|=1} \inf_t\{t \mid \mathrm{E}[\exp(\langle Z, z\rangle)^2/t^2] \leq 2\}.$$

1. $Z = [Z_1, Z_2, \dots, Z_n]^\top \in \mathbb{R}^{n\times p}$ の各行 $Z_i \in \mathbb{R}^p$ が独立な等方的サブガウシアン確率変数とする.
2. ある半正定対称行列 $\Sigma \in \mathbb{R}^{p\times p}$ を用いて $X = Z\Sigma$ であるとする.

Theorem (Rudelson and Zhou (2013))

$\|Z_i\|_{\psi_2} \leq \kappa$ $(\forall i)$ とする．ある普遍定数 c_0 が存在し $m = c_0\frac{\max_i(\Sigma_{i,i})^2}{\phi_{\mathrm{RE}}^2(k,9,\Sigma)}$ に対し，$n \geq 4c_0m\kappa^4\log(60ep/(m\kappa))$ ならば

$$P\left(\phi_{\mathrm{RE}}(k, 3, \widehat{\Sigma}) \geq \frac{1}{2}\phi_{\mathrm{RE}}(k, 9, \Sigma)\right) \geq 1 - 2\exp(-n/(4c_0\kappa^4)).$$

つまり，**真の分散共分散行列が制限固有値条件を満たす**
⇒ **経験的分散共分散行列も高い確率で同条件を満たす.**

凸最適化を用いないスパース推定法の性質

- 情報量規準型推定量: Massart (2003), Bunea et al. (2007), Rigollet and Tsybakov (2011).

$$\min_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 + C\sigma^2\|\beta\|_0 \left\{1 + \log\left(\frac{p}{\|\beta\|_0}\right)\right\}.$$

- Bayes 推定量: Dalalyan and Tsybakov (2008), Alquier and Lounici (2011), Suzuki (2012).

オラクル不等式: X に 何も条件を課さずに 次の不等式が成り立つ:

$$\frac{1}{n}\|X\beta^* - X\hat{\beta}\|^2 \leq C\sigma^2\frac{d}{n}\log\left(1 + \frac{p}{d}\right).$$

- ミニマックス最適.
- 凸正則化推定法と大きなギャップ.
- 計算量と統計的性質のトレードオフ.

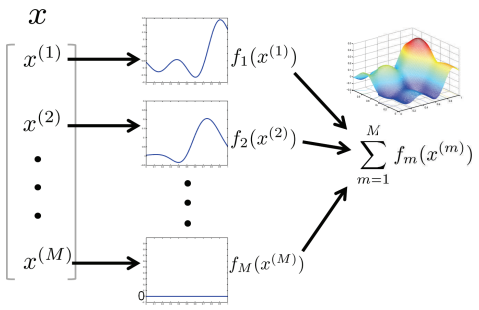
より高度なトピック：線形回帰以外のスパース推定

複数のデータソースを取捨選択し統合する方法.

- ノンパラメトリックスパース推定：スパース加法モデル
- 高次相関の推定：低ランクテンソルモデル

スパース加法モデル

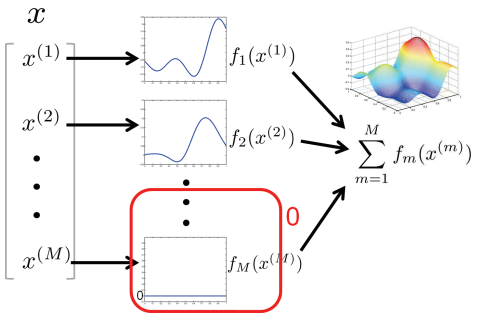
線形モデルから非線形モデルへ拡張 (高次元ノンパラメトリック推定)



$$y_i = \sum_{j=1}^{p} x_i^{(j)} b_j + \epsilon_i \quad \longrightarrow \quad y_i = \sum_{j=1}^{p} f_j(x_i^{(j)}) + \epsilon_i$$

スパース加法モデル

線形モデルから非線形モデルへ拡張 (高次元ノンパラメトリック推定)



$$y_i = \sum_{j=1}^{p} x_i^{(j)} b_j + \epsilon_i \quad \longrightarrow \quad y_i = \sum_{j=1}^{p} f_j(x_i^{(j)}) + \epsilon_i$$

Multiple Kernel Learning

Multiple Kernel Learning (MKL) (Lanckriet et al., 2004, Bach et al., 2004)

$$\hat{f} = \sum_{m=1}^{M} \hat{f}_m \leftarrow \min_{f_m\in\mathcal{H}_m} \sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M} f_m(x_i)\right)^2 + C\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}$$

($\mathcal{H}_m$: ある再生核ヒルベルト空間)

- グループ正則化の無限次元拡張
- スパースな解: 多くの $\hat{f}_m$ が 0.
- 有限次元最適化問題に帰着 (表現定理) (Sonnenburg et al., 2006, Rakotomamonjy et al., 2008, Suzuki and Tomioka, 2009)

Multiple Kernel Learning

Multiple Kernel Learning (MKL) (Lanckriet et al., 2004, Bach et al., 2004)

$$\hat{f} = \sum_{m=1}^{M} \hat{f}_m \leftarrow \min_{f_m\in\mathcal{H}_m} \sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M} f_m(x_i)\right)^2 + C\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}$$

($\mathcal{H}_m$: ある再生核ヒルベルト空間)

- グループ正則化の無限次元拡張
- スパースな解: 多くの $\hat{f}_m$ が 0.
- 有限次元最適化問題に帰着 (表現定理) (Sonnenburg et al., 2006, Rakotomamonjy et al., 2008, Suzuki and Tomioka, 2009)

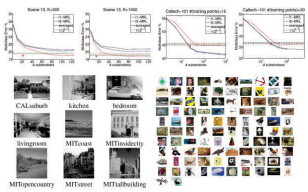
正則化項は一般化してよい

$$\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m} \quad \longrightarrow \quad \psi((\|f_m\|_{\mathcal{H}_m})_{m=1}^{M}),$$

e.g., ℓ_p-正則化 (Micchelli and Pontil, 2005, Kloft et al., 2009), エラスティックネット正則化 (Shawe-Taylor, 2008, Tomioka and Suzuki, 2009), Variable Sparsity Kernel Learning (VSKL) (Aflalo et al., 2011).

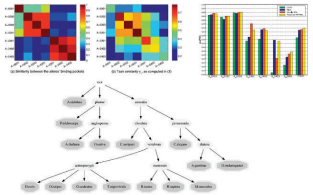
応用例

- コンピュータビジョン



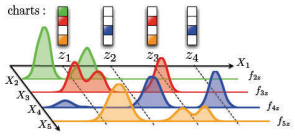
(Gehler&Nowozin, CVPR2009)

- バイオインフォマティクス



(Widmer et al., BMC Bioinformatics 2010)

- Time varying coefficient model



$f_{(t)}(x) = \beta_{(t)}^\top x$ (Lu et al., 2015)

スパース加法モデルの種々の統計的性質

様々な手法・状況における収束レート.

- 一般の正則化型学習 (Suzuki, 2011)

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*} \|f^*\|_\psi)^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right).$$

- 真がスパースな場合の elastic-net 型正則化学習 (Suzuki and Sugiyama, 2013)

(L1) $$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} + \frac{d\log(M)}{n}\right),$$

(Elastic) $$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d\log(M)}{n}\right),$$

- ベイズ法: ガウス過程回帰 + モデル平均 (Suzuki, 2012)

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^\circ\|_n^2\right] = \mathcal{O}_p\left[\sum_{m\in I_0} n^{-\frac{1}{1+s_m}} + \frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)\right].$$

→ Restricted Eigenvalue Condition なしでミニマクスレートを達成.

ノンパラメトリックな状況でも真の非ゼロ要素数が支配的.

スペクトル条件 (s)

$0 < s < 1$: 再生核ヒルベルト空間の複雑さ.

カーネルの展開 (cf., Mercer の定理):

$$k_m(x, x') = \sum_{\ell=1}^\infty \mu_{\ell,m}\phi_{\ell,m}(x)\phi_{\ell,m}(x'),$$

ただし $\{\phi_{\ell,m}\}_{\ell=1}^\infty$ は $L_2(P)$ に関する正規直交基底.

スペクトル条件 (s)

ある $0 < s < 1$ が存在して

$$\mu_{\ell,m} \le C\ell^{-\frac{1}{s}} \quad (\forall \ell, m).$$

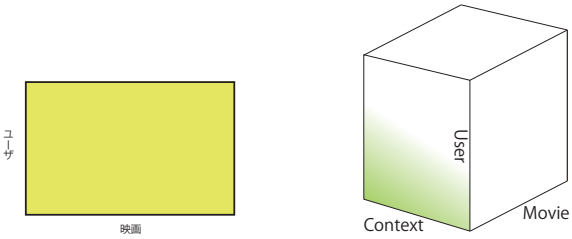
- 大きな s は複雑，小さな s は単純.

複雑さ s の再生核ヒルベルト空間における最適な学習レート: $O_p(n^{-\frac{1}{1+s}})$.

Proposition (Steinwart et al. (2009))

$$\mu_{\ell,m} \sim \ell^{-\frac{1}{s}} \Leftrightarrow \log N(B(\mathcal{H}_m), \epsilon, L_2(P)) \sim \epsilon^{-2s}$$

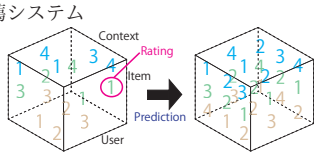
テンソルデータ



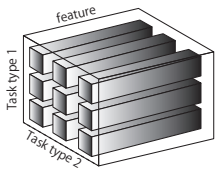
行列 $X_{ij} = \sum_{r=1}^d u_{r,i}^{(1)} u_{r,j}^{(2)}$ $\longrightarrow$ テンソル (多重線形オペレーター) $X_{ijk} = \sum_{r=1}^d u_{r,i}^{(1)} u_{r,j}^{(2)} u_{r,k}^{(3)}$

その他応用

- テンソル補完：推薦システム



- マルチタスク学習



- 自然言語処理 (単語のベクトル表現)
- 時空間解析
- 関係データ

テンソルの回帰問題

$$Y_i = \langle X_i, \mathcal{A}^* \rangle + W_i.$$

$\mathcal{A}^*, X_i \in \mathbb{R}^{M_1 \times \dots M_K}$: テンソル.

$\langle X_i, \mathcal{A}^* \rangle := \sum_{j_1,\dots,j_K} X_{i,(j_1,\dots,j_K)} \mathcal{A}^*_{j_1,\dots,j_K}$.

$W_i \sim \mathcal{N}(0, \sigma^2)$: observational noise.

E.g., $X_i = e_{j_1} \otimes e_{j_2} \otimes \cdots \otimes e_{j_K}$ ならテンソル補完問題.

仮定: $\mathcal{A}^*$ は “低ランク”.

テンソルの回帰問題

$$Y_i = \langle X_i, \mathcal{A}^* \rangle + W_i.$$

$\mathcal{A}^*, X_i \in \mathbb{R}^{M_1 \times \dots M_K}$: テンソル.
$\langle X_i, \mathcal{A}^* \rangle := \sum_{j_1,\dots j_K} X_{i,(j_1,\dots j_K)} \mathcal{A}^*_{j_1,\dots j_K}$.
$W_i \sim \mathcal{N}(0, \sigma^2)$: observational noise.
E.g., $X_i = e_{j_1} \otimes e_{j_2} \otimes \cdots \otimes e_{j_K}$ ならテンソル補完問題.

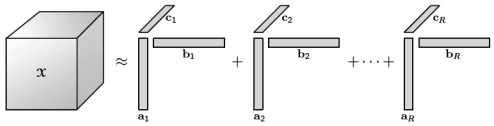
仮定: $\mathcal{A}^*$ は "低ランク".

凸正則化のアプローチ:

$$\min_{\mathcal{A} \in \mathbb{R}^{M_1 \times M_2 \times \cdots \times M_K}} \sum_{i=1}^{n} (Y_i - \langle X_i, \mathcal{A}^* \rangle)^2 + \mathrm{pen}(\mathcal{A}).$$

- CP-ランクの凸拡張
- 和型 Schatten-1 ノルム (Tomioka et al., 2011)
- 畳み込み型 Schatten-1 ノルム (Tomioka and Suzuki, 2013)

テンソルのランク: CP-ランク



CP-分解 (Canonical Polyadic Decomp.)
(Hitchcock, 1927a,b)

(figure is from (Kolda and Bader, 2009))

$\mathcal{X}_{ijk} = \sum_{r=1}^{d} a_{ir} b_{jr} c_{kr} =: [[A, B, C]]$.
CP-分解は CP-ランクを定義する.

- CP-分解は **NP-困難**.
- CP-ランクはテンソルの辺の長さより大きくなりうる.
- 直交分解が存在するとは限らない (対称テンソルでも).
- 凸拡張の計算は NP-困難. 行列のランクの凸拡張は計算しやすい.

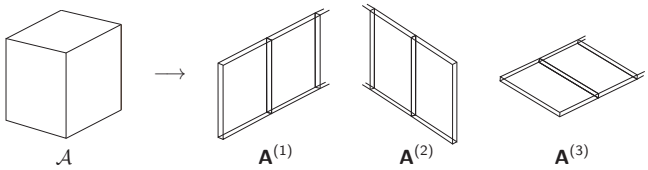
和型 Schatten-1 ノルム

$$\|\mathcal{A}\|_{S_1/1} := \sum_{k=1}^{K} \|\mathbf{A}^{(k)}\|_{\mathrm{Tr}}$$

和型 Schatten-1 ノルム正則化

$$\widehat{\mathcal{A}} = \arg\min_{\mathcal{A}} \|\mathcal{Y} - \mathcal{A}\|_F^2 + \lambda_n \|\mathcal{A}\|_{S_1/1}.$$

Tucker-ランクが小さなテンソルを推定.



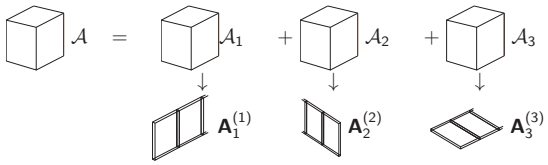
畳み込み型 Schatten-1 ノルム

$$\|\mathcal{A}\|_{\underline{S_1/1}} := \inf_{\mathcal{A}=\mathcal{A}_1+\mathcal{A}_2+\cdots+\mathcal{A}_K} \sum_{k=1}^{K} \|\mathbf{A}_k^{(k)}\|_{\mathrm{Tr}}$$

畳み込み型 Schatten1 ノルム正則化

$$\widehat{\mathcal{A}} = \arg\min_{\mathcal{A}} \quad \|\mathcal{Y} - \mathcal{A}\|_F^2 + \lambda_n \|\mathcal{A}\|_{\underline{S_1/1}},$$
$$\text{s.t.} \quad \mathcal{A} = \sum_{k=1}^{K} \mathcal{A}_k, \|\mathbf{A}_k^{(k')}\|_{S_\infty} \le \frac{\alpha}{K} \sqrt{N/n_{k'}} \ (\forall k' \neq k).$$

ランクの小さな方向を見つける.



収束レート解析

$$\min_{\mathcal{A} \in \mathbb{R}^{M_1 \times M_2 \times \cdots \times M_K}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, \mathcal{A} \rangle)^2 + \mathrm{pen}(\mathcal{A}).$$

- 和型 (Tomioka et al., 2011)
$$\frac{1}{N}\|\widehat{\mathcal{A}} - \mathcal{A}^*\|_F^2 \le \frac{C}{n} \left(\frac{1}{K} \sum_{k=1}^{K} (\sqrt{M_k} + \sqrt{N/M_k}) \right)^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k} \right)^2$$
- 畳み込み形 (Tomioka and Suzuki, 2013):
$$\frac{1}{N}\|\widehat{\mathcal{A}} - \mathcal{A}^*\|_F^2 \le \frac{C}{n} \left(\max_k (\sqrt{M_k} + \sqrt{N/M_k}) \right)^2 \min_k r_k$$
- **Square deal** (Mu et al., 2014):
$$\frac{1}{M}\|\hat{A} - A^*\|_2^2 \le C \frac{\min\{\prod_{k \in I_1} r_k, \prod_{k \in I_2} r_k\}}{n} \left(\prod_{k \in I_1} M_k + \prod_{k \in I_2} M_k \right),$$
ただし, I_1 と I_2 はインデックス $\{1, \dots, K\}$ の任意の分割.

計算量: 少ない.
統計的性質: 改善可能. → ベイズ推定法

ベイズ推定

ベイズ法で低ランクテンソル推定を行った時の統計的性質を調べる.

- より少ない条件で**最適レート**を達成.

ベイズ事後分布の収束

$\|\mathcal{A}^*\|_{\max,2} \le R\sigma_{\mathrm{p}}$ を仮定.
(真のテンソルは事前分布の台に含まれている)

Theorem

ある $n, \{M_k\}_k$ とは関係ない定数 C が存在して,

(固定デザイン)
$$\mathrm{E}_{Y_{1:n}|X_{1:n}}\left[\frac{1}{2\sigma^2}\int \|\mathcal{A}-\mathcal{A}^*\|_n^2 \mathrm{d}\Pi(\mathcal{A}|X_{1:n}, Y_{1:n})\right] \le C\frac{d(\sum_{k=1}^K M_k)}{n}(1\vee R^2)\log\left(K\frac{\sigma_{\mathrm{p}}^K}{\xi}\sqrt{nR^K}\right).$$

(ランダムデザイン)
$$\mathrm{E}_{Y_{1:n}|X_{1:n}}\left[\frac{1}{2\sigma^2}\int \|\mathcal{A}-\mathcal{A}^*\|_{L_2}^2 \mathrm{d}\Pi(\mathcal{A}|X_{1:n}, Y_{1:n})\right] \le C\frac{d(\sum_{k=1}^K M_k)}{n}(1\vee R^{2(K+1)})\log\left(K\frac{\sigma_{\mathrm{p}}^K}{\xi}\sqrt{nR^K}\right).$$

このレートはデザイン行列に 強凸性を仮定せず に得られる.

ミニマックス最適性

ミニマックス最適リスク：どんな推定量 でも超えられないリスク.

$$\mathcal{H}_d(R) := \left\{\mathcal{A}\in\mathbb{R}^{M_1\times\cdots\times M_K} \mid \text{CP-ランク } d, \|\mathcal{A}\|_{\max,2}\le R\right\}.$$

(ランク d のテンソルの集合)

Theorem

$\mathcal{H}_d(R)$ 上のテンソル推定量のミニマックス最適レートは以下で与えられる:

$$\min_{\widehat{\mathcal{A}}}\max_{\mathcal{A}^*\in\mathcal{H}_d(R)} \mathrm{E}[\|\widehat{\mathcal{A}}-\mathcal{A}^*\|_{L_2}^2] \gtrsim \frac{d(M_1+\cdots+M_K)}{n}.$$

※ 先の収束レートは log 項を除いてミニマックス最適レートを達成.

凸正則化手法との比較

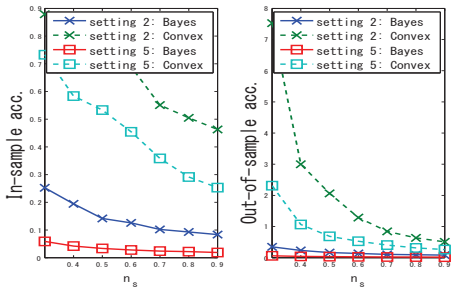


Figure: ベイズ推定法と凸正則化法 (和型 Schatten-1 ノルム) との比較.

誤差のスケール

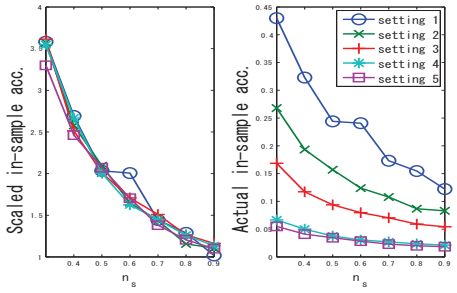


Figure: 固定デザイン: The scaled predictive accuracy (left) and the actual predictive accuracy (right) against the number of samples.

$$\text{scaled accuracy} = \frac{\text{actual accuracy}}{d(\sum_{k=1}^K M_k)/n}$$

Outline

1. スパース推定のモデル
2. いろいろなスパース正則化
3. スパース推定の理論
   - $n \gg p$ の理論
   - $n \ll p$ の理論
   - より高度なトピック
4. 高次元線形回帰の検定
5. スパース推定の最適化手法

バイアス除去による方法

アイディア：Lasso 推定量 $\hat{\beta}$ からバイアスを除去.
(van de Geer et al., 2014, Javanmard and Montanari, 2014)

$$\tilde{\beta} = \hat{\beta} + MX^\top(Y - X\hat{\beta})$$

M が $(X^\top X)^{-1}$ なら,

$$\tilde{\beta} = \beta^* + (X^\top X)^{-1}X^\top\epsilon.$$

→ バイアスなし, (漸近) 正規.

問題点: $p \gg n$ のとき, $X^\top X$ は非可逆.

M の求め方:

$$\min_{M\in\mathbb{R}^{p\times p}} \quad |\hat{\Sigma}M^\top - I|_\infty.$$

($|\cdot|_\infty$ は中身をベクトルとみなした無限大ノルム)

Theorem (Javanmard and Montanari (2014))

$\epsilon_i \sim N(0,\sigma^2)$ $(i.i.d.)$ とする.

$$\sqrt{n}(\tilde{\beta}-\beta^*) = Z + \Delta, \qquad Z \sim N(\mathbf{0}, \sigma^2 M\hat{\Sigma}M^\top), \quad \Delta = \sqrt{n}(M\hat{\Sigma}-I)(\beta^*-\hat{\beta}).$$

また, X がランダムで分散共分散行列が正定な時, $\lambda_n = c\sigma\sqrt{\log(p)/n}$ とすると,

$$\|\Delta\|_\infty = O_p\left(\frac{d\log(p)}{\sqrt{n}}\right).$$

※ $n \gg d^2\log^2(p)$ ならば $\Delta \approx 0$ で, $\sqrt{n}(\tilde{\beta}-\beta^*)$ はほぼ正規分布に従う.
→ 信頼区間の構築や検定ができる.

M の求め方:

$$\min_{M\in\mathbb{R}^{p\times p}} \quad |\hat{\Sigma}M^\top - I|_\infty.$$

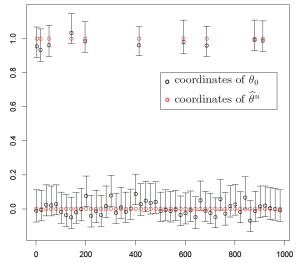
($|\cdot|_\infty$ は中身をベクトルとみなした無限大ノルム)

Theorem (Javanmard and Montanari (2014))

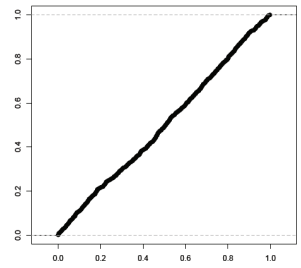
$$\sqrt{n}(\tilde{\beta}-\beta^*) = \underbrace{Z}_{\text{正規分布}} + \underbrace{\Delta}_{\substack{X \text{ が非可逆であることによる残りカス} \\ \text{条件が良い時 } 0 \text{ へ収束}}}$$

※ $n \gg d^2\log^2(p)$ ならば $\Delta \approx 0$ で, $\sqrt{n}(\tilde{\beta}-\beta^*)$ はほぼ正規分布に従う.
→ 信頼区間の構築や検定ができる.

数値実験



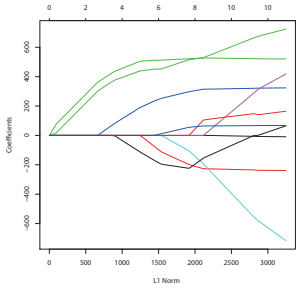
(a) 人工データにおける 95%信頼区間. $(n,p,d) = (1000,600,10)$.



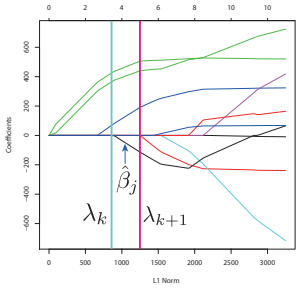
(b) 人工データにおける p 値の CDF. $(n,p,d) = (1000,600,10)$.

図は Javanmard and Montanari (2014) から引用.

共分散検定統計量 (Lockhart et al., 2014)



共分散検定統計量 (Lockhart et al., 2014)



$J = \mathrm{supp}(\hat{\beta}(\lambda_k))$, $J^* = \mathrm{supp}(\beta^*)$,
$\tilde{\beta}(\lambda_{k+1}) := \underset{\beta:\beta_J\in\mathbb{R}^{|J|},\beta_{J^c}=0}{\mathrm{argmin}} \quad \|Y - X_J\beta_J\|^2 + \lambda_{k+1}\|\beta_J\|_1.$
$J^* \subseteq J$ ならば (つまり $\beta^*_j = 0$),

$$T_k = \left(\langle Y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle Y, X\tilde{\beta}(\lambda_{k+1})\rangle\right)/\sigma^2 \xrightarrow{\mathrm{d}} \mathrm{Exp}(1) \quad (n,p\to\infty).$$

Outline

1. スパース推定のモデル
2. いろいろなスパース正則化
3. スパース推定の理論
   - $n \gg p$ の理論
   - $n \ll p$ の理論
   - より高度なトピック
4. 高次元線形回帰の検定
5. スパース推定の最適化手法

スパース推定における最適化の問題意識

$$R(\beta) = \underbrace{\sum_{i=1}^{n} \ell(y_i, x_i^\top \beta)}_{f(\beta):\text{ロス関数}} + \underbrace{\psi(\beta)}_{\text{正則化項}}$$
$$= f(\beta) + \psi(\beta)$$

- ψ が尖っている→微分不可能.
- 尖っている関数の最適化は 難しい.
- f はなめらかな場合が多い.
- ψ の構造を利用 すれば，あたかも R がなめらか であるかのように最適化可能.

スパース推定における最適化の問題意識

$$R(\beta) = \underbrace{\sum_{i=1}^{n} \ell(y_i, x_i^\top \beta)}_{f(\beta):\text{ロス関数}} + \underbrace{\psi(\beta)}_{\text{正則化項}}$$
$$= f(\beta) + \psi(\beta)$$

- ψ が尖っている→微分不可能.
- 尖っている関数の最適化は 難しい.
- f はなめらかな場合が多い.
- ψ の構造を利用 すれば，あたかも R がなめらか であるかのように最適化可能.

典型例：$L1$ 正則化 $\psi(\beta) = C\sum_{j=1}^{p} |\beta_j|$ → 座標ごとに分かれている.

一次元の最適化 $\min_b\{(b-y)^2 + C|b|\}$ は 簡単.

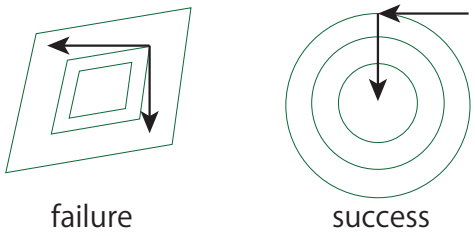
座標降下法

座標降下法の手順

1. 座標 $j \in \{1, \ldots, p\}$ を 何らかの方法 で選択.
2. j 番目の座標 β_j を更新.（以下は更新方法の例）
   - $\beta_j^{(k+1)} \leftarrow \mathrm{argmin}_{\beta_j} R([\beta_1^{(k)}, \ldots, \beta_j, \ldots, \beta_p^{(k)}])$.
   - $g_j = \frac{\partial f(\beta^{(k)})}{\partial \beta_j}$ として，
     $\beta_j^{(k+1)} \leftarrow \mathrm{argmin}_{\beta_j} \ \langle g_j, \beta_j \rangle + \psi_j(\beta_j) + \frac{\eta_k}{2}\|\beta_j - \beta_j^{(k)}\|^2$.

- 座標を一つずつではなく複数個選ぶことも多い: ブロック座標降下法.
- 座標はあるルールに従って選んだり，ランダムに選んだりする.

座標降下の注意点



- 左側: 座標降下が失敗．降下方向がない.
- 座標降下を成功させるには，目的関数は降下方向がなくてはいけない．理想的には座標ごとに分離 $f(x) = \sum_{j=1}^{p} f_j(x_j)$.

座標降下法の収束

$$\min_x\{P(x)\} = \min_x\{f(x) + \psi(x)\} = \min_x\{f(x) + \textstyle\sum_{j=1}^{p} \psi_j(x_j)\}.$$

仮定: f は 各座標ごと に γ-平滑 $(|\partial_{x_j} f(x) - \partial_{x_j} f(x + ae_j)| \leq \gamma a)$

- サイクリックな更新 (Saha and Tewari, 2013, Beck and Tetruashvili, 2013)

$$P(x^{(t)}) - R(x^*) \leq \frac{\gamma p \|x^{(0)} - x^*\|^2}{2t} = O(p\gamma/t).$$

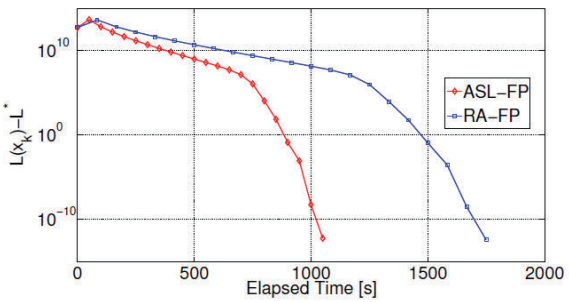
- ランダムな選択 (Nesterov, 2012, Richtárik and Takáč, 2014)
  - 加速なし: $O(p\gamma/t)$.
  - Nesterov の加速: $O(\gamma(p/t)^2)$ (Fercoq and Richtárik, 2013).
  - f が α-強凸: $O(\exp(-C(\alpha/\gamma)t/p))$.
  - f が α-強凸 + 加速: $O(\exp(-C\sqrt{\alpha/\gamma}t/p))$ (Lin et al., 2014).

座標はランダムに選択すれば良い.

有用なサーベイ論文: Wright (2015).

大規模データにおける座標降下法

Hydra: 並列分散計算を用いた座標降下法 (Richtárik and Takáč, 2013, Fercoq et al., 2014).



大規模 Lasso ($p = 5 \times 10^8$, $n = 10^9$) における Hydra の計算効率 (Richtárik and Takáč, 2013)．128 ノード，4,096 コア.

近接勾配法型手法

$$\underbrace{f(\beta)}_{\text{線形近似}} + \psi(\beta)$$

$g_k \in \partial f(\beta^{(k)})$, $\bar{g}_k = \frac{1}{k}\sum_{\tau=1}^{k} g_\tau$.

- 近接勾配法:

$$\beta^{(k+1)} = \arg\min_{\beta\in\mathbb{R}^p}\left\{ g_k^\top \beta + \psi(\beta) + \frac{\eta_k}{2}\|\beta - \beta^{(k)}\|^2 \right\}.$$

- 正則化双対平均法 (Xiao, 2009, Nesterov, 2009):

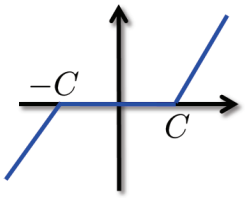
$$\beta^{(k+1)} = \arg\min_{\beta\in\mathbb{R}^p}\left\{ \bar{g}_k^\top \beta + \psi(\beta) + \frac{\eta_k}{2}\|\beta\|^2 \right\}.$$

鍵となる計算は近接写像: $\mathrm{prox}(\mathbf{q}|\psi) := \arg\min_{\mathbf{x}}\left\{ \psi(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{q}\|^2 \right\}$.
$L1$ 正則化なら簡単に計算できる (Soft-thresholding 関数).
ψ が劣モジュラ関数の Lovasz 拡張：劣モジュラ性を使った効率的最適化

近接写像の例: L_1 正則化

$$\begin{aligned}\mathrm{prox}(\mathbf{q}|C\|\cdot\|_1) &= \arg\min_{\mathbf{x}}\left\{ C\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x}-\mathbf{q}\|^2 \right\} \\ &= (\mathrm{sign}(q_j)\max(|q_j| - C, 0))_j.\end{aligned}$$

→ Soft-thresholding 関数. 解析解!



近接勾配法型手法の収束レート

f の性質	滑らか	非滑らか
強凸	$\exp(-\sqrt{\alpha/L}k)$	$\frac{1}{k}$
非強凸	$\frac{1}{k^2}$	$\frac{1}{\sqrt{k}}$

- 滑らかな場合，Nesterov の加速法を使った時の収束レートを示している (Nesterov, 2007, Zhang et al., 2010).
- 加速法を使わなければ，それぞれ $\exp(-(\alpha/L)k)$, $\frac{1}{k}$ になる.
- 上のオーダーは勾配情報のみを用いる方法 (First order method) の中で最適.

拡張ラグランジアン型手法

$$\begin{aligned}&\min_{\beta} f(\beta) + \psi(\beta) \\ \Leftrightarrow\quad &\min_{x,y} f(x) + \psi(y) \ \text{ s.t. } \ x = y.\end{aligned}$$

最適化の難しさを 分離.

拡張ラグランジアン: $\mathcal{L}(x,y,\lambda) = f(x) + \psi(y) - \lambda^\top(y-x) + \frac{\rho}{2}\|y-x\|^2$.

乗数法 (Hestenes, 1969, Powell, 1969, Rockafellar, 1976)

1. $(x^{(k+1)}, y^{(k+1)}) = \mathrm{argmin}_{x,y}\,\mathcal{L}(x,y,\lambda^{(k)})$.
2. $\lambda^{(k+1)} = \lambda^{(k)} - \rho(y^{(k+1)} - x^{(k+1)})$.

x, y での同時最適化はやや面倒 → 交互方向乗数法.

交互方向乗数法

$$\min_{x,y} f(x) + \psi(y) \ \text{ s.t. } \ x = y.$$

$$\mathcal{L}(x,y,\lambda) = f(x) + \psi(y) - \lambda^\top(y-x) + \frac{\rho}{2}\|y-x\|^2.$$

交互方向乗数法 (Gabay and Mercier, 1976)

$$x^{(k+1)} = \arg\min_{x} f(x) + \lambda^{(k)\top}x + \frac{\rho}{2}\|y^{(k)} - x\|^2$$

$$y^{(k+1)} = \arg\min_{y} \psi(y) - \lambda^{(k)\top}y + \frac{\rho}{2}\|y - x^{(k+1)}\|^2 (= \mathrm{prox}(x^{(k+1)} + \lambda^{(k)}/\rho|\psi/\rho))$$

$$\lambda^{(k+1)} = \lambda^{(k)} - \rho(y^{(k+1)} - x^{(k+1)})$$

- x, y の同時最適化は交互に最適化することで回避.
- y の更新は近接写像. L_1 正則化のような場合は簡単.
- 構造的正則化への拡張も容易.
- 最適解に収束する保証あり.
- 一般的には $O(1/k)$ (He and Yuan, 2012), 強凸ならば線形収束 (Deng and Yin, 2012, Hong and Luo, 2012).

確率的最適化

サンプル数の多い大規模データ で有用.
一回の更新にすべてのサンプルを読み込まないでも大丈夫.

- オンライン型
  - FOBOS (Duchi and Singer, 2009)
  - RDA (Xiao, 2009)
- バッチ型
  - SVRG (Stochastic Variance Reduced Gradient) (Johnson and Zhang, 2013)
  - SDCA (Stochastic Dual Coordinate Ascent) (Shalev-Shwartz and Zhang, 2013)
  - SAG (Stochastic Averaging Gradient) (Le Roux et al., 2013)

- 確率的交互方向乗数法: Suzuki (2013), Ouyang et al. (2013), Suzuki (2014).

確率的最適化の基本方針

$$R(w) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(z_i, w)}_{\text{重い}} + \psi(w)$$

大きな問題を分割して個別に処理

- データ $\{z_i\}_{i=1}^n$ から一つ（もしくは）少数をランダムに抽出
- パラメータ w を更新

※各更新で $O(1)$ (もしくは $O(p)$) の計算量.

- オンライン型：SGD, SDA
  - 一般: $O(1/\sqrt{T})$
  - 強凸: $O(1/T)$
- バッチ型：SVRG, SAG, SDCA
  - $\exp(-\frac{T}{n+\lambda/\gamma})$

オンライン型手法

$$\min_w \mathrm{E}_{Z\sim P(Z)}[\ell(Z, w)] + \psi(w)$$

- ランダムに $z_t \sim P(Z)$ を生成.
- $g_t \in \partial_w \ell(z_t, w^{(t-1)})$, $\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$. ※勾配は 1 つのデータのみで計算
- SGD (Stochastic Gradient Descent):

$$w^{(t)} = \arg\min_{w\in\mathbb{R}^p}\left\{g_t^\top w + \psi(w) + \frac{1}{2\eta_t}\|w - w^{(t-1)}\|^2\right\}.$$

- SDA (Stochastic Dual Averaging):

$$w^{(t)} = \arg\min_{w\in\mathbb{R}^p}\left\{\bar{g}_t^\top w + \tilde{\psi}(w) + \frac{1}{2\eta_t}\|w\|^2\right\}.$$

[仮定] $\mathrm{E}[\|g_t\|^2] \le G^2$, $\mathrm{E}[\|x_t - x^*\|^2] \le D^2$ $(\forall t)$ (SDA は二つ目の仮定を緩められる)

- $R(\bar{w}^{(T)}) \le \frac{2GR}{\sqrt{T}}$ (非平滑, 非強凸) Polyak-Ruppert 平均化 $\bar{w}^{(T)} = \frac{\sum_t w^{(t)}}{T}$
- $R(\bar{w}^{(T)}) \le \frac{2G^2}{\mu T}$ (非平滑, 強凸) 多項式平均化 $\bar{w}^{(T)} = \frac{\sum_t t w^{(t)}}{T(T+1)/2}$

バッチ型確率的最適化手法

サンプルサイズは**固定** (オンライン型は次から次へとデータを読み込む)

$$\frac{1}{n}\sum_{i=1}^{n}\ell(z_i, w) + \psi(w)$$

代表的な 3 つの手法

- **確率的平均勾配法**
  Stochastic Average Gradient descent, SAG
  (Le Roux et al., 2012, Schmidt et al., 2013, Defazio et al., 2014)
- **確率的分散縮小勾配法**
  Stochastic Variance Reduced Gradient descent, SVRG
  (Johnson and Zhang, 2013, Xiao and Zhang, 2014)
- **確率的双対座標降下法**
  Stochastic Dual Coordinate Ascent, SDCA
  (Shalev-Shwartz and Zhang, 2013)

- SAG と SVRG は 主問題 を解く方法.
- SDCA は 双対問題 を解く方法.

問題設定

$$P(x) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell_i(x)}_{\text{平滑}} + \underbrace{\psi(x)}_{\text{強凸}}$$

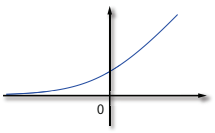
仮定:

- ℓ_i: 損失関数は γ-平滑. $(\|\nabla\ell_i(x) - \nabla\ell_i(x')\| \le L\|x - x'\|)$
- ψ: 正則化関数は λ-強凸. $(\psi(x) - \frac{\lambda}{2}\|x\|^2$ が凸関数)
  典型的には $\lambda = O(1/n)$ または $O(1/\sqrt{n})$.

例:

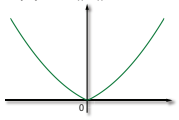
損失関数

- 平滑化ヒンジ損失
- ロジスティック損失



正則化関数

- L_2 正則化
- エラスティックネット正則化
- $\tilde{\psi}(x) + \lambda\|x\|^2$



バッチ型手法の特色

- 1 回の更新に 1 データのみ利用 (オンライン型と同じ),
- 線形収束 (オンライン型と違う):

$$T > (n + \gamma/\lambda)\log(1/\epsilon)$$

回の更新で ϵ 誤差を達成.
ただし γ-平滑な損失と λ-強凸な正則化を仮定.

- 通常の勾配法 (各反復で全データ使用) の計算量:

$$T > n\gamma/\lambda\log(1/\epsilon)$$

[通常] $n(\gamma/\lambda)\log(1/\epsilon)$ → [確率的] $(n + \gamma/\lambda)\log(1/\epsilon)$

確率的双対座標降下法

双対問題 において確率的座標降下法を適用.

- 双対の各座標は各データ点に対応.
  → 「一つの座標を更新」 = 「一つのデータ点を観測して更新」
- 誤差 ϵ までに必要な反復数:

$$T \ge C\left(n + \frac{\gamma}{\lambda}\right)\log\left(\frac{n + \gamma/\lambda}{\epsilon}\right).$$

→ 線形収束.

- 二重ループは必要ない.
  ※他のバッチ型手法である SVRG は二重ループ

仮定:

- $\ell(z_i, \cdot)$ は γ-平滑 ($\leftrightarrow$ $\ell^*(z_i, \cdot)$ は $1/\gamma$-強凸)
- ψ は λ-強凸 ($\leftrightarrow$ ψ^* は $1/\lambda$-平滑)

ルジャンドル変換

双対空間 (勾配の空間) で定義された関数.

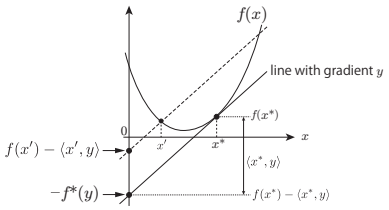
Definition (ルジャンドル変換)

(非凸かもしれない) 関数 $f : \mathbb{R}^p \to \bar{\mathbb{R}}$ に対し，その共役関数は次のように定義される:

$$f^*(y) := \sup_{x \in \mathbb{R}^p} \{\langle x, y \rangle - f(x)\}.$$

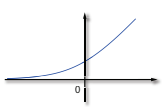
f から f^* への変換をルジャンドル変換と呼ぶ.

※ 共役関数は凸関数.　※ 凸関数 f に対し f^* はその 裏の姿.

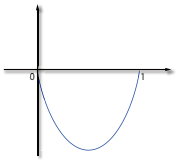


例

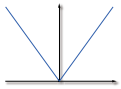
	$f(x)$	$f^*(y)$
二乗損失	$\frac{1}{2}x^2$	$\frac{1}{2}y^2$
ヒンジ損失	$\max\{1-x, 0\}$	$\begin{cases} y & (-1 \le y \le 0), \\ \infty & (\text{otherwise}). \end{cases}$
ロジスティック損失	$\log(1+\exp(-x))$	$\begin{cases} (-y)\log(-y) + (1+y)\log(1+y) & (-1 \le y \le 0), \\ \infty & (\text{otherwise}). \end{cases}$
L_1 正則化	$\|x\|_1$	$\begin{cases} 0 & (\max_j \lvert y_j \rvert \le 1), \\ \infty & (\text{otherwise}). \end{cases}$
L_p 正則化 $(p > 1)$	$\sum_{j=1}^d \lvert x_j \rvert^p$	$\sum_{j=1}^d \frac{p-1}{p^{\frac{p}{p-1}}} \lvert y_j \rvert^{\frac{p}{p-1}}$



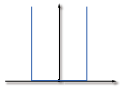
ロジスティック



その双対



L_1 ノルム



その双対

双対問題

$\exists f_i : \mathbb{R} \to \mathbb{R}$ なる関数があって $\ell(z_i, x) = f_i(a_i^\top x)$ と書けると仮定.
$A = [a_1, \ldots, a_n]$ とする.

(Primal)
$$\inf_{x \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n f_i(a_i^\top x) \; + \; \psi(x) \right\}$$

[Fenchel の双対定理]

$$\inf_{x \in \mathbb{R}^p} \{f(A^\top x) + n\psi(x)\} = - \inf_{y \in \mathbb{R}^n} \{f^*(y) + n\psi^*(-Ay/n)\}$$

(Dual)
$$\inf_{y \in \mathbb{R}^n} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n f_i^*(y_i)}_{\text{強凸}} \; + \; \underbrace{\psi^*\left(-\frac{1}{n}Ay\right)}_{\text{平滑}} \right\}$$

なお次の式を用いた:

- $f(\alpha) = \sum_{i=1}^n f_i(\alpha_i)$ に対し $f^*(\beta) = \sum_{i=1}^n f_i^*(\beta_i)$.
- $\tilde{\psi}(x) = n\psi(x)$ に対し $\tilde{\psi}^*(y) = n\psi^*(y/n)$.

確率的双対座標降下法

確率的双対座標降下法 (Shalev-Shwartz and Zhang, 2013)

Iterate the following for $t = 1, 2, \ldots$

- Pick up an index $i \in \{1, \ldots, n\}$ uniformly at random.
- Calculate $x^{(t-1)} = \nabla\psi^*(-A^\top y^{(t-1)}/n)$.
- Update the i-th coordinate y_i so that the objective function is decreased:
  - $$y_i^{(t)} \in \operatorname*{argmin}_{y_i \in \mathbb{R}} \left\{ f_i^*(y_i) - \langle x^{(t-1)}, a_i y_i \rangle + \frac{1}{2\eta} \|y_i - y_i^{(t-1)}\|^2 \right\}$$
    $$= \operatorname{prox}(y_i^{(t-1)} + \eta a_i^\top x^{(t-1)} | \eta f_i^*),$$
  - $y_j^{(t)} = y_j^{(t-1)}$ (for $j \neq i$).

- $x^{(t)}$ は主問題の解になる
- 一回ごとの計算量は小さい

バッチ型確率的最適化手法のまとめ

各種手法の性質

Method	SDCA	SVRG	SAGA
主/双対	双対	主	主
メモリ効率	✓	✓	△
その他	$\ell_i(\beta) = f_i(x_i^\top \beta)$	二重ループ	平滑な正則化

誤差 ϵ までの計算量:

$$\left(n + \frac{\gamma}{\lambda}\right) \log(1/\epsilon).$$

加速法もある (Catalyst (Lin et al., 2015), Acc-SDCA (Lin et al., 2014)):

$$\left(n + \sqrt{\frac{n\gamma}{\lambda}}\right) \log(1/\epsilon)$$

確率的交互方向乗数法

線形制約付き双対問題

Let $A = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{p \times n}$.

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^n f_i(a_i^\top w) + \psi(B^\top w) \right\} \qquad \text{(Primal)}$$

$$\Leftrightarrow \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n f_i^*(x_i) + \psi^*\left(\frac{y}{n}\right) \;\middle|\; Ax + By = 0 \right\} \qquad \text{(Dual)}$$

近接写像の計算しやすい ψ と 線形変換 $B^\top$ の組みで多くの正則化項をカバー：構造的正則化.

確率的交互方向乗数法 = 確率的最適化 + 交互方向乗数法

- オンライン型: Suzuki (2013), Ouyang et al. (2013)
- バッチ型 (双対座標降下): SDCA-ADMM (Suzuki, 2014)

$$T = O\left(\left(n + \sqrt{\frac{n\gamma}{\lambda}}\right) \log(1/\epsilon)\right).$$

→ 加速法と同じ収束レート.

構造的正則化の例

- 重複あり のグループ正則化
- Fused Lasso 正則化 (Tibshirani et al., 2005, Jacob et al., 2009).
- 低ランクテンソル正則化 (Signoretto et al., 2010; Tomioka et al., 2011).
- Robust PCA (Candés et al., 2009).

どれも近接写像の計算しやすい関数 ψ と行列 B を用いて

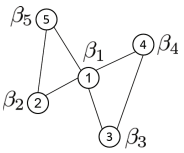
$$\psi(B^\top x) = \tilde{\psi}(x)$$

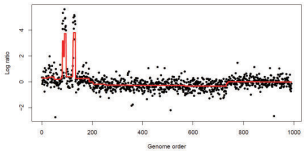
のように書ける.

構造的正則化の例: (一般化) Fused Lasso

$$\psi(\beta) = C \sum_{(i,j)\in E} |\beta_i - \beta_j|.$$

(Tibshirani et al. (2005), Jacob et al. (2009))





Fused lasso による遺伝子データ解析 (Tibshirani and Taylor '11)

TV デノイジング (Chambolle '04)

確率的交互方向乗数法

確率的座標降下型交互方向乗数法 (Suzuki, 2014)

Split the index set $\{1,\dots,n\}$ into K groups $(I_1, I_2, \dots, I_K)$.
For each $t = 1, 2, \dots$
  Choose $k \in \{1,\dots,K\}$ uniformly at random, and set $I = I_k$.

$$\begin{aligned}
y^{(t)} &\leftarrow \arg\min_y \Big\{ n\psi^*(y/n) - \langle w^{(t-1)}, Ax^{(t-1)} + By \rangle \\
&\qquad + \frac{\rho}{2}\|Ax^{(t-1)} + By\|^2 + \frac{1}{2}\|y - y^{(t-1)}\|_Q^2 \Big\}, \\
x_I^{(t)} &\leftarrow \arg\min_{x_I} \Big\{ \textstyle\sum_{i\in I} f_i^*(x_i) - \langle w^{(t-1)}, A_I x_I + By^{(t)} \rangle \\
&\qquad + \frac{\rho}{2}\|A_I x_I + A_{\setminus I} x_{\setminus I}^{(t-1)} + By^{(t)}\|^2 + \frac{1}{2}\|x_I - x_I^{(t-1)}\|_{G_{I,I}}^2 \Big\}, \\
w^{(t)} &\leftarrow w^{(t-1)} - \gamma\rho\{ n(Ax^{(t)} + By^{(t)}) \\
&\qquad - (n - n/K)(Ax^{(t-1)} + By^{(t-1)}) \}.
\end{aligned}$$

where Q, G are some appropriate positive semidefinite matrices.

単純化した方法

$$Q = \rho(\eta_B \mathrm{I}_d - B^\top B), \ G_{I,I} = \rho(\eta_{Z,I} \mathrm{I}_{|I|} - Z_I^\top Z_I),$$

とすれば, $\mathrm{prox}(q|\psi) + \mathrm{prox}(q|\psi^*) = q$ を用いて更新式を単純化できる.

単純化した SDCA-ADMM

For $q^{(t)} = y^{(t-1)} + \frac{B^\top}{\rho\eta_B}\{w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t-1)})\}$, let

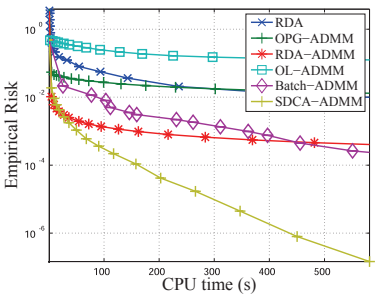
$$y^{(t)} \leftarrow q^{(t)} - \mathrm{prox}(q^{(t)}|n\psi(\rho\eta_B \ \cdot \ )/(\rho\eta_B)),$$

For $p_I^{(t)} = x_I^{(t-1)} + \frac{Z_I^\top}{\rho\eta_{Z,I}}\{w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t)})\}$, let

$$x_i^{(t)} \leftarrow \mathrm{prox}(p_i^{(t)}|f_i^*/(\rho\eta_{Z,I})) \quad (\forall i \in I).$$

★ x の更新は並列化可能.
y の更新は ψ に関する近接写像 (計算しやすい).

確率的交互方向乗数法



まとめ

- 様々なスパースモデリング
  - $L1$ 正則化
  - グループ正則化
  - トレースノルム正則化
- Lasso の漸近的振る舞い
  - 漸近分布
  - Adaptive Lasso のオラクルプロパティ
  - 制限固有値条件→ $\|\hat{\beta} - \beta^*\|^2 = O_p(d \log(p)/n)$
- より進んだ話題
  - スパース加法モデル
  - 低ランクテンソル推定
- 検定
  - バイアス除去法, 共分散検定統計量
- 最適化手法
  - 座標降下法
  - 近接勾配法
  - (交互方向) 乗数法
  - 確率的最適化手法→大規模データにおける最適化で有用

J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. S. Nath, and S. Raman. Variable sparsity kernel learning. *Journal of Machine Learning Research*, 12:565–592, 2011.

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

T. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22: 327–351, 1951.

A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In Y. S. J.C. Platt, D. Koller and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 25–32, Cambridge, MA, 2008. MIT Press.

F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *the 21st International Conference on Machine Learning*, pages 41–48, 2004.

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop 2007*, 2007.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer, 2011.

F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

G. R. Burket. *A study of reduced-rank models for multiple prediction*, volume 12 of *Psychometric monographs*. Psychometric Society, 1964.

E. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, 346: 589–592, 2008.

E. Candès and T. Tao. The power of convex relaxations: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56:2053–2080, 2009.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52 (12):5406–5425, 2006.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM TR12-14, 2012.

D. Donoho. Compressed sensing. *IEEE Transactions of Information Theory*, 52 (4):1289–1306, 2006.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2873–2908, 2009.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 2001.

O. Fercoq and P. Richtárik. Accelerated, parallel and proximal coordinate descent. Technical report, 2013. arXiv:1312.5799.

O. Fercoq, Z. Qu, P. Richtárik, and M. Takáč. Fast distributed coordinate descent for non-strongly convex losses. In *Proceedings of MLSP2014: IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers & Mathematics with Applications*, 2:17–40, 1976.

T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall Ltd, 1999.

B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numerical Analisis*, 50(2):700–709, 2012.

M. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory & Applications*, 4:303–320, 1969.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6:164–189, 1927a.

F. L. Hitchcock. Multilple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7:39–79, 1927b.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. Technical report, 2012. arXiv:1208.3922.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, pages 248–264, 1975.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning*, page to appear, 2014.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-va pdf`.

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate ℓ_p-norm multiple kernel learning. In *Advances in Neural Information Processing Systems 22*, pages 997–1005, Cambridge, MA, 2009. MIT Press.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

N. Le Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc., 2012.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems 25*, 2013.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. Technical report, 2015. arXiv:1506.02186.

Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. Technical report, 2014. arXiv:1407.1296.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

K. Lounici, A. Tsybakov, M. Pontil, and S. van de Geer. Taking advantage of sparsity in multi-task learning. 2009.

J. Lu, M. Kolar, and H. Liu. Post-regularization confidence bands for high dimensional nonparametric models with local sparsity, 2015. arXiv:1503.02978.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

N. Meinshausen and P. B uhlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the 31th International Conference on Machine Learning*, pages 73–81, 2014.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming, Series B*, 120:221–259, 2009.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

H. Ouyang, N. He, L. Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

M. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, London, New York, 1969.

A. Rakotomamonjy, F. Bach, S. Canu, and G. Y. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

G. Raskutti and M. J. Wainwright. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:389–427, 2012.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71(5):1009–1030, 2009.

P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. Technical report, 2013. arXiv:1310.2059.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144:1–38, 2014.

P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1: 97–116, 1976.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions of Information Theory*, 39, 2013.

A. Saha and A. Tewari. On the non-asymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013.

M. Schmidt, N. Le Roux, and F. R. Bach. Minimizing finite sums with the stochastic average gradient, 2013. hal-00860051.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14: 567–599, 2013.

J. Shawe-Taylor. Kernel learning for novelty detection. In *NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, Whistler, 2008.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

N. Srebro, N. Alon, and T. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances in Neural Information Processing Systems (NIPS) 17*, 2005.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pages 79–93, 2009.

T. Suzuki. Unifying framework for fast learning rate of non-sparse multiple kernel learning. In *Advances in Neural Information Processing Systems 24*, pages 1575–1583, 2011. NIPS2011.

T. Suzuki. Pac-bayesian bound for gaussian process regression and multiple kernel additive model. In *JMLR Workshop and Conference Proceedings*, volume 23, pages 8.1–8.20, 2012. Conference on Learning Theory (COLT2012).

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning*, pages 392–400, 2013.

T. Suzuki. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *Proceedings of the 31th International Conference on Machine Learning*, pages 736–744, 2014.

T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *The Annals of Statistics*, 41(3): 1381–1405, 2013.

T. Suzuki and R. Tomioka. SpicyMKL, 2009. arXiv:0909.5026.

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

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. 67(1):91–108, 2005.

R. Tomioka and T. Suzuki. Sparsity-accuracy trade-off in MKL. In *NIPS 2009 Workshop: Understanding Multiple Kernel Learning Methods*, Whistler, 2009.

R. Tomioka and T. Suzuki. Convex tensor decomposition via structured schatten norm regularization. In *Advances in Neural Information Processing Systems 26*, page accepted, 2013. NIPS2013.

R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems 24*, pages 972–980, 2011. NIPS2011.

S. van de Geer, P. B uehlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems 23*, 2009.

L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24:2057–2075, 2014.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statist*, 38(2):894–942, 2010.

P. Zhang, A. Saha, and S. V. N. Vishwanathan. Regularized risk minimization by nesterov's accelerated gradient methods: Algorithmic extensions and empirical studies. *CoRR*, abs/1011.0472, 2010.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5):2109–2144, 2009.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

田中利幸. 圧縮センシングの数理. *IEICE Fundamentals Review*, 4(1):39–47, 2010.