スパースコーディングの研究動向 *

笠井 裕之 †

2014 年 2 月

Abstract

スパースコーディング（Sparse Coding: SC）は，入力信号を少数の基底ベクトルの重み付き線形和で表現する技術の一般クラスとして捉えることができる．具体的には，基底ベクトル $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m \in \mathbb{R}^n$ と重み係数 $\boldsymbol{x} \in \mathbb{R}^m$ を用いて，各入力信号 $\boldsymbol{b} \in \mathbb{R}^n$ を $\boldsymbol{b} \approx \sum_i \boldsymbol{a}_i x_i$ を満たすように表現する．基底ベクトルの数 m は入力信号次元 n よりも大きな過完備基底 $(n < m)$ であるため，非常に多くの表現パターンがある．元々は，視覚野中の受容野のニューロンを効果的に符号化するための計算モデルとして提案された [1] が，その後，推薦システムや画像や音響信号などのメディア信号処理，イベント検知や DNA 解析など多数の分野に応用され，例えば画像識別の分野で最先端の性能を示している．

本技術分野は比較的若く，最初の重要な発表は Mallat と Zhang らにより 1993 年に発表された Matching Pursuit [2] であり，それ以後の貪欲法へとつながる．第二の発表は Chen と Donoho，Saunders により 1995 年に発表された Basis Pursuit [3] である．これは ℓ_1 ノルムによりスパース性を評価することで，スパース解の導出を凸計画問題としてとらえたものである．これら二つの発表を皮切りに，より深いアルゴリズム解析やアプリケーション開発が進められた．特に，2001 年に発表された Donoho らの研究成果 [**?**] は，この分野で後に重要な問いかけとなる，Pursuit 法が正解を導くことへの保証性やその条件に対する問題に対して，部分的ではあるがその答えを示すものであった．本分野は，信号処理と応用数学との交差点に位置するため，近似理論や調和解析の応用数学者，科学者，統計学者，またコンピュータサイエンスや電気電子学，地球物理学などの様々な分野のエンジニアが研究に参加している．

本稿では，このようなスパースコーディングの基礎からその応用までの技術の一部について紹介する．特に，一般的にスパース解の求解手法は計算負荷が高く，大規模データを対象とした場合はそれが顕著であることから，それらに対応すべく提案された様々な高効率な最適化手法の最新動向についても紹介する．

1 問題定義と解のユニーク性 [4]

1.1 (P_0) 問題と (P_1) 問題の定義

$\boldsymbol{x} \in \mathbb{R}^m$ と $n < m$ を満たすフルランクの行列 $\mathbf{A} \in \mathbb{R}^{n \times m}$ に対して Underdetermined な線形システム $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ を考える．本問題は無限の解を持つため，解の妥当性を評価する関数として $J(\boldsymbol{x})$ を導入し一般化問題 (P_J) を“$(P_J) : \min_{\boldsymbol{x}} J(\boldsymbol{x})$ subject to $\boldsymbol{b} = A\boldsymbol{x}$”として定義する．ここで厳密な凸関数 $J(\cdot)$ を選択すれば唯一の解が得られ，例えば，ℓ_2 ノルムとして (P_2) 問題を定義すると，ヘッセ行列の正定値性により $\nabla^2 \|\boldsymbol{x}\|_2^2 = 2\mathbf{I} \succeq \mathbf{0}$ であり凸であるから，唯一の解

*本資料は，情報処理学会研究報告オーディオビジュアル複合情報処理（AVM）2014-AVM-84(8), 1-10, 2014-02-14 として発表した資料に加筆したものである．

†Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan (kasai@is.uec.ac.jp).

$\hat{\boldsymbol{x}} = \mathbf{A}^{\dagger}\boldsymbol{b} = \mathbf{A}^{T}(\mathbf{A}\mathbf{A}^{T})^{-1}\boldsymbol{b}$ を得る．しかし閉形式を持たない場合でも凸性を有していればユニークな解を得ることができるため，大域的最小解への収束を保証する最適化手法を適用することで解を得られる．一方，本稿が目標とするスパース性に対しては，$\boldsymbol{x}$ の少数の非零係数しか存在しない場合，つまり l_0 ノルムを $\|\boldsymbol{x}\|_0 = \#i : x_i \neq 0$ と定義し $\|\boldsymbol{x}\|_0 \ll m$ である場合，$\boldsymbol{x}$ は“スパース”であると定義することで，(P_0) 問題を以下に定義する．

$$(P_0): \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{b} = \mathbf{A}\boldsymbol{x} \tag{1}$$

(P_2) 問題の解はユニークであるが，(P_0) 問題の場合は，ℓ_0 ノルムは離散・非連続性を有するため標準的な凸解析手法は適用できない．従って，(i)『解はユニークであるか? またどのような状況の場合に?』，(ii)『もし解がある場合，その解が大域的最小解であることを簡単に検証することは可能か?』という問いが残る．

一方，(P_0) 問題は，組み合わせ探索問題（離散最適化問題）であり，その複雑度は m の指数乗で増加するため NP-hard 問題となる．そこで，連続関数で近似し，特に凸問題を考えることで，最適化を容易にする．$0 < p < 1$ の場合は非凸関数となるが，$p \geq 1$ の場合，凸関数となり，解の唯性が保証され，数理的な研究成果による汎用解法が利用できる．一方，ℓ_0 の解と大きく異なることも回避し，且つスパースな解を持つ必要があるため，$\|\boldsymbol{x}\|_1$ を ℓ_1 ノルム $\sum_i |x_i|$ として定義し，(P_1) 問題を考える．

$$(P_1): \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{b} = \mathbf{A}\boldsymbol{x} \tag{2}$$

この問題は *Basis Pursuit* [3] と呼ばれ，凸最適化問題であり，インコヒーレントな列ベクトルから構成される $\mathbf{A}$ に対して，(P_0) 問題が十分にスパースな解を持つ場合は常に，その解はユニークであり，且つ (P_1) 問題の解と一致することが知られている [5]．ここで (P_1) 問題は，線形計画法に帰着でき標準的な最適化解法により得ることができる．

1.2 解のユニーク性： スパークと相互コヒーレンス

行列 $\mathbf{A}$ の零空間 (null-space) は，$\mathbf{A}\boldsymbol{x} = \mathbf{0}$ を満たすベクトルから構成される空間であり，$\ker(\mathbf{A}) = \{\boldsymbol{x} \in \mathbb{R}^m : \mathbf{A}\boldsymbol{x} = \mathbf{0}\}$ で定義される．この零空間を用いて，$\boldsymbol{b} = \mathbf{A}\boldsymbol{x}$ を満たすような疎ベクトル $\boldsymbol{x}$ の解の一意性を示す方法のうち，Donoho らにより 2003 年に提案されたスパーク (*spark*) を用いる方法がある [6]．スパークは，行列 $\mathbf{A}$ の一次従属な列ベクトルの数の最小値を指し，$spark(\mathbf{A}) = \min_{\boldsymbol{z} \in \ker(\mathbf{A})\{\backslash \mathbf{0}\}} \|\boldsymbol{z}\|_0$ で定義される．ここで零空間内にあるベクトル $\boldsymbol{x}$ は，スパークの定義から $\|\boldsymbol{x}\|_0 \geq spark(\mathbf{A})$ を満たす．スパークはスパース解のユニーク性を考える上で重要であり Rao らにより 1998 年に提案され，心理統計学の研究でも Kruskal Rank [7] としても提案された．さて，以下にスパークを用いた解のユニーク性についての定理を示す．

Theorem 1.1 (Uniqueness - Spark)**.** 線形方程式 $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}$ が $\|\boldsymbol{x}\|_0 < spark(\mathbf{A})/2$ を満たす $\boldsymbol{x}$ を解として持つならば，その解は必ず最もスパースな解である．

この定理は $spark(\mathbf{A}) > 2K$ ならば $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}$ を満たす K-スパースなベクトル $\boldsymbol{x}$ が多くともひとつ存在することを意味する．より大きなスパーク値には意味があり，どの程度大きな値なのか問題となる．定義から，スパークは $2 \leq spark(\mathbf{A}) \leq n+1$ であり，例えば $\mathbf{A}$ がランダム i.i.d のガウス分布から構成される場合，確率 1 で $spark(\mathbf{A}) = n+1$，つまり一次従属な n 個の列がないことが知られている．

一方，スパークを求めるには行列 $\mathbf{A}$ の列の全て組合せを探索する必要があるため厳密に評価することは極めて難しいことから，相互コヒーレンス (mutual-coherence) が導入された．これは，与えられた行列 $\mathbf{A}$ の異なる列間の最大の絶対値正規化内積を示し，行列 $\mathbf{A}$ の k 番目列

を $\boldsymbol{a}_k$ と定義すると，$\mu(\mathbf{A}) = \max_{1\le i,j\le m, i\neq j} \frac{|\boldsymbol{a}_i^T \boldsymbol{a}_j|}{\|\boldsymbol{a}_i\|_2 \cdot \|\boldsymbol{a}_j\|_2}$. と定義される．相互コヒーレンスは，行列 $\mathbf{A}$ を構成する列同士の依存関係を特徴づける指標である．ユニタリー行列の場合は，各列ペアで直交しているため相互コヒーレンスは零である．一方，列数が行数よりも多い一般行列 $(n < m)$ においては，ユニタリー行列の特性に出来るだけ近い小さな値を期待する．Donoho と Huo は $n \times m$ のランダム直交行列については，インコヒーレント（互いに依存しない）であり，$\mu(\mathbf{A}_{n,m})$ は典型的に $\sqrt{\log(nm)/n}$ に比例することを示している [8]．前述の通り，相互コヒーレンスは比較的計算が容易であり，計算困難なスパークの下界を与える．以上から，いかなる行列 $\mathbf{A} \in \mathbb{R}^{n\times m}$ に対しても，$spark(\mathbf{A}) \ge 1 + 1/\mu(\mathbf{A})$ が成り立つという補題が示されることから，ユニーク性についての定理を，以下のように相互コヒーレンスから得る．

Theorem 1.2 (Ubiqueness - Mutual Coherence)**.** $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}$ が $\|\boldsymbol{x}\|_0 < (1/2)(1 + 1/\mu(\mathbf{A}))$ を満たす $\boldsymbol{x}$ を解として持つならば，その解は最もスパースな解である．

ここで，**Theorem 1.1**と **Theorem 1.2** とでは仮定が異なり，前者は後者より強力な定理である．相互コヒーレンスは，単にスパークの下界を示しているに過ぎない．$\mu(\mathbf{A})$ は $1/\sqrt{n}$ より小さくならないため，**Theorem 1.2**の Cardinality 境界は $\sqrt{n}/2$ より大きくならない．一方，スパークは前述の通り，容易に n と同じ程度の大きさとなるため，**Theorem 1.1**は $n/2$ の境界を与える．$n = 100$ の場合は，相互コヒーレンスにおけるスパース数最大値は 5 であり，スパークでは 50 であることからも分かる．

1.3 最大事後確率推定との関係

ここでスパース解の導出問題と最大事後確率 (MAP) 推定との関係について言及しておく．標準的な生成モデルでは，復元誤差 $\boldsymbol{b} - \sum_i \boldsymbol{a}_i x_i$ の分布 $p(\boldsymbol{b}|\mathbf{A}, \boldsymbol{x})$ を，分散共分散行列 $\sigma^2\mathbf{I}$ を持つゼ零平均のガウス分布 $\mathcal{N}(0, \sigma^2\mathbf{I})$ で表現する．ここで，係数 $\boldsymbol{x}$ の事前分布 $\pi(\boldsymbol{x})$ をスパース性を考慮して $\pi(x_i) \propto \exp(-\beta\phi(x_i))$ のラプラス分布で表現するとする．尚，$\phi(\cdot)$ はスパース関数であり $\phi(x_j) = \|x_i\|_1$ を考える．ここで，n 次元の入力信号 $(b_1, \cdots, b_n)^T$ と対応する未知の m 次元の係数 $(x_1, \cdots, x_m)^T$ の学習データ k 個を考える．事後確率はベイズの定理から $p(\mathbf{A}, \boldsymbol{x}|\boldsymbol{b}) \propto p(\boldsymbol{b}|\mathbf{A}, \boldsymbol{x})\pi(\boldsymbol{x})\pi(\mathbf{A})$ であるから，最大事後確率 (MAP) 推定により，以下の最適化問題を解くことで最適な $\boldsymbol{a}_i$ と x_i を得ることになる．但し，基底ベクトル $\boldsymbol{a}_i \in \mathbb{R}^n$ の事前確率は一様分布とした．

$$\min_{\boldsymbol{a}_i, \boldsymbol{x}} \sum_{j=1}^{k} \frac{1}{2\sigma^2} \|\boldsymbol{b}^{(j)} - \sum_{i=1}^{m} \boldsymbol{a}_i x_i^{(j)}\|^2 + \beta \sum_{j=1}^{k}\sum_{i=1}^{m} \|x_i^{(j)}\|_1 \tag{3}$$

ここで各列が $\boldsymbol{b}$ で構成される入力信号行列を $\mathbf{B} \in \mathbb{R}^{n\times k}$，各列が $\boldsymbol{x}$ から構成される係数行列を $\mathbf{X} \in \mathbb{R}^{m\times k}$ とし，また Frobenius ノルムを $\|\cdot\|_F$ と表すと，式 (3) は “ $\min_{\mathbf{A},\mathbf{X}}(1/2\sigma^2)\|\mathbf{B} - \mathbf{A}\mathbf{X}\|_F^2 + \beta\sum_{j,i}\|\mathbf{X}_{j,i}\|_1$ ” となる．これは後述の式 (6) の (Q_1^λ) 問題と一致することが分かる．

1.4 様々なスパース正則化項

(P_0) 問題及び (P_1) 問題の制約付き最適化問題は，ラグランジュ乗数を導入して制約無し最適化問題として定義される．例えば **2.2**で紹介する (Q_1^λ) 問題は，(P_1^ϵ) 近似問題の制約条件を損失項とし，目的関数を正則化項として問題を書き換えている．そのような観点から (P_0) 及び (P_1) 問題は，ℓ_0 ノルム及び ℓ_1 ノルムを正則化項として持つ制約無し問題と等価に扱える．特に，ℓ_1 ノルムを正則化項として推定する方法は，統計分野で Lasso (Least Absolute Shrinkage and Selection Operator) [9] と呼ばれる．一方，その他にもスパース性を評価する正則化が多数

提案されており，グループ $\ell_{1,2}$ ノルム $\sum_{j=1}^{k}\sqrt{\sum_{i\in G_j} x_i^2}$ を用いる Group Lasso が提案されている [10]．G_j は j 番目グループを示し，行列のグループ（例えば行あるいは列）単位でスパース性を評価する手法である．例えばマルチタスク学習やアレイ解析で共通して変数を求める場合に利用される．その他 Fused Lasso は 1 次元の順序関係がある特徴量において隣接特徴量間の変化を $\sum_{i=1}^{n-1}|x_{i+1}-x_i|$ として評価する．この 2 次元版が Total Variation で画像処理で使用される．さらに ℓ_1 ノルムと ℓ_2 ノルムの重み付き和を用いた Elastic Net はリッジ回帰と Lasso の中間的立場をとる．

2 求解手法と性能保証 [4]

2.1 厳密解法 Pursuit 法と性能保証

2.1.1 Pursuit 法

(P_0) 問題及び (P_1) 問題の解法には，ℓ_0 ノルムを直接求める“貪欲法”や，ℓ_p $(p\in(0,1)$) ノルム等の滑らかな関数を導入して ℓ_0 ノルムを連続あるいは滑らかな近似により置き換える“凸緩和法”，また確率推論ベースの手法等がある．本稿では前者 2 つの手法について紹介する．

貪欲法 (Greedy Pursuit)　貪欲法は，(P_0) 問題を解くことを目的として，全ての組合せを調べず,『入力信号 $\boldsymbol{b}$ との残差誤差 $\boldsymbol{r}$ と相関の高い列ベクトル $\boldsymbol{a}_i$ は（スパース）係数に含まれる』という仮説の下，次々に係数を選択していく手法である．本手法は統計モデリングにおいて *forward stepwise regression* と呼ばれ，1960 年代から広く利用されてきたものである．ここでは，最も代表的な Orthogonal Matching Pursuit (OMP) 法について説明する．最初に，残差 $\boldsymbol{r}_0=\boldsymbol{b}$，求めるべき係数のインデックス集合（サポート: Support）$\mathcal{S}_0=\emptyset$，として初期化する．そして計算処理ループでは，まず Sweep ステップで全ての $j(\forall 1\le j\le m)$ について $\epsilon(j)=\min_{x_j}\|\boldsymbol{r}_{k-1}-x_j\boldsymbol{a}_j\|^2$ を計算し，次の Update Support ステップで $j_0=\arg\min_{i\notin\mathcal{S}_{k-1}}\{\epsilon(j)\}$ を求め，サポートを $\mathcal{S}_k=\mathcal{S}_{k-1}\cup j_0$ により更新する．次の Estimate ステップで，$\boldsymbol{x}_k=\arg\min_{\boldsymbol{y}}\|\boldsymbol{b}-\mathbf{A}_{\mathcal{S}_k}\boldsymbol{y}\|_2$ より，これまでに選択された列を用いて最も良い似近となる係数 $\boldsymbol{x}_k$ を求める．ここで，$\mathbf{A}_{\mathcal{S}_k}$ はサポート $\mathcal{S}_k$ のインデックスに対応する $n\times|\mathcal{S}_k|$ サイズの部分行列である．次の Update Residual ステップでは，$\boldsymbol{r}_k=\boldsymbol{b}-\mathbf{A}_{\mathcal{S}_k}\boldsymbol{x}_k$ により残差を更新する．これを残差誤差が閾値 ϵ_0 以下になる等の終了条件まで行い，最後に $\mathcal{S}_k$ に対応する係数が非零でそれ以外が零の $\boldsymbol{x}$ を出力し処理を終了する．k_0 を係数の個数とすると，全探索の計算量が $\mathcal{O}(nm^{k_0}k_0^2)$ から $\mathcal{O}(nmk_0)$ に削減される．

一方で，OMP 法には数多くの変形があり，例えば，Estimate ステップでサポート全体に対して解き直さない簡易な Matching Pursuit (MP) 法や，Sweep ステップで全てを評価せず条件を満たした段階で中止する Weak-MP 法がある．さらに，最も大きな k 個の内積値を持つサポートだけを用いるアイデアで，サポート評価は最初 1 度だけ行ない，これに基づいて係数を算出する Thresholding 法もある．

凸緩和法 (Convex Relaxation)　ℓ_p ノルム緩和方法として，Gorodnitsky らは 1997 年に FOCUSS (FOCal Underdetermined System Solver) 法 [11] を提案した．これは反復再重み付け最小二乗法 (Iterative Reweighted Least Squares: IRLS) を用いて ℓ_p ノルムの局所最小解を探索する方法であり，ℓ_p ノルムを扱いながら重み付き ℓ_2 ノルムの計算として解く．本手法は，ある固定値に収束することが保証されているが，必ずしもそれは大域解ではなく (P_0) 問題の大域的最小解の条件が不明である．

一方，凸近似が可能な ℓ_1 ノルムで置き換える方法がある．この場合 (P_1) 問題は線形計画問題として帰着でき，内点法やシンプレックス法，Homotopy 法等で解くことで大域解を得るこ

とができ，貪欲法と比べて洗練された方法といえる．尚，様々なソルバーは Web 上に公開されている．

2.1.2 Pursuit 法の性能保証

我々の関心事は『$\boldsymbol{b} = \mathbf{A}\boldsymbol{x}$ が $\|\boldsymbol{x}\|_0 = k_0$ のスパース解をもち，$k_0 < spark(\mathbf{A})/2$ を満たすことを仮定する場合，貪欲法や凸緩和法がそのスパース解を復元できるかどうか?』である．全ての k_0 且つ全ての行列 $\mathbf{A}$ に期待できないが，十分にスパースな解を実際に持つ場合，(P_0) 問題に対して解を復元することが保証される．具体的には，**2.1.1**で述べた OMP 法については，$\|\boldsymbol{x}\|_0 < (1/2)(1+1/\mu(\mathbf{A}))$ に従う解 $\boldsymbol{x}$ が存在する時，閾値 $\epsilon_0 = 0$ を条件として正確にその解を得ることを保証する．一方，**2.1.1**で述べた凸緩和による Basis Pursuit では，$\|\boldsymbol{x}\|_0 < (1/2)(1+1/\mu(\mathbf{A}))$ に従う解 $\boldsymbol{x}$ が存在する時，$\boldsymbol{x}$ は (P_1) 問題の唯一の解で，且つ (P_0) 問題の唯一の解となる．

上記の定理は，(P_0) 問題の解を近似するために貪欲法と凸緩和法を使用する動機付けになるため重要であるが，これらの結果は弱く，次元 n の内 $\sqrt{n}$ より少ない極めてスパースな場合でのみ保証するものである．これは，あらゆる信号と係数密度を対象とした最悪の場合の分析結果であるからである．しかし実際には，相互コヒーレンスはスパース性が求められる部分的な状況を示しているに過ぎず，過去の数多くの実験から，ランダム行列 $\mathbf{A}$ の場合，貪欲法と凸緩和法の性能は，前記の限界を破る状況下においても良い性能を示すことが知られている．

2.2 近似解法と安定性

2.2.1 厳密解から近似解 $(P_0), (P_1)$ から $(P_0^\epsilon), (P_1^\epsilon)$

(P_0) 問題の設定は一般的に非現実的であることから，厳密制約 $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ を緩和して，ペナルティ関数 $\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ を用いて，(P_0^ϵ) 問題として評価されることが多い．

$$(P_0^\epsilon): \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}\|_2 \le \epsilon \tag{4}$$

(P_0) 問題と (P_0^ϵ) 問題を同じ問題に適用した場合，(P_0^ϵ) 問題が (P_0) 問題より少ない非零係数を持つこともある．ここで (P_0^ϵ) 問題をノイズ除去問題としてとらえると，十分にスパースなベクトル $\boldsymbol{x}_0$ が $\|\boldsymbol{z}\|_2^2 = \epsilon^2$ を用いて $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}_0 + \boldsymbol{z}$ を満たす場合，(P_0) 問題がノイズの無いデータ $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}_0$ の解を求めるのと同様に，(P_0^ϵ) 問題は $\boldsymbol{x}_0$ を求めていると言える．但し，この場合，解のユニーク性の議論は当てはまらず，解の安定性として評価される．

2.2.2 近似解の安定性

(P_0^ϵ) 問題の解の近似手法を述べる前に，基本的な質問に答える必要がある．それは『ノイズが混在した観測信号 $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}_0 + \boldsymbol{e}$ を得るものとし，$\boldsymbol{x}_0$ を近似して $\|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}\|_2 \le \epsilon$ の下で $\boldsymbol{x}_0^\epsilon = \min_{\boldsymbol{x}}\|\boldsymbol{x}\|_0$ を解くことで解 $\boldsymbol{x}_0^\epsilon$ を得る場合，この近似はどの程度良いのか?』である．これは (P_0) 問題で議論した“解のユニーク性”に対する自然な拡張である．

これに対する一つの回答として，Restricted Isometry Property (RIP) に基づく評価方法が示されている．ℓ_2 正規化された列から成る $\mathbf{A} \in \mathbb{R}^{n\times m}(n < m)$ と，スカラー整数値 $s \ge n$ に対して，行列 $\mathbf{A}$ からの s 列からなる部分行列 $\mathbf{A}_s$ を考える．$\delta_s(< 1)$ を任意の $\boldsymbol{c} \in \mathbb{R}^s$ について，$(1-\delta_s)\|\boldsymbol{c}\|_2^2 \le \|\mathbf{A}_s\boldsymbol{c}\|_2^2 \le (1+\delta_s)\|\boldsymbol{c}\|_2^2$ を満たす最小値と定義する．これは，行列 $\mathbf{A}$ から得られるいかなる s 列のサブセット行列は，エネルギーをほとんど失わないあるいは増加しない直交変換のように振る舞うことを意味する．ここで，$\boldsymbol{x}_0 \in \mathbb{R}^m$ が (P_0^ϵ) の実行可能な解であり，$\mathbf{A}$ が $\delta_{2s_0} < 1$ を満たす $2s_0$ の RIP 特性に従う場合を考える．この時，許容誤差 ϵ $(i.e., \|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}_0\|_2 \le \epsilon)$ を含む $\boldsymbol{b}$ を与えるものとする．このとき，(P_0^ϵ) の全ての解 $\boldsymbol{x}_0^\epsilon$ は，$\|\boldsymbol{x}_0^\epsilon - \boldsymbol{x}_0\|_2^2 < \frac{4\epsilon^2}{1-\mu(\mathbf{A})(2\|\boldsymbol{x}_0\|_0 - 1)}$ を満たす．

2.2.3 Pursuit 法の拡張

前述の Pursuit 法は誤差を許容する場合にも適応可能である．OMP 法に代表される貪欲法では，アルゴリズム中の停止 (収束) 条件を $\epsilon_0 = \epsilon$ とすることで，制約条件 $\|\boldsymbol{b} - \mathrm{A}\boldsymbol{x}\|_2 \leq \epsilon$ が満たされるまで解ベクトル中の非零数を増やしていけば良い．同様にして，ℓ_1 への凸緩和法の場合においては，以下の (P_1^ϵ) 問題を定義して解くことになる．

$$(P_1^\epsilon): \qquad \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}\|_2 \leq \epsilon \tag{5}$$

本問題は Basis pursuit denoising (BPDN) と呼ばれ，内点法をはじめとする多くの既存解法が適用可能であり，Web から入手可能な様々な最適化パッケージツールが利用できる．しかし大規模問題を扱う場合，そのような一般目的の二次計画最適化ツールの速度は遅いことが知られている．

一方，適切なラグランジュ乗数 λ を導入し，(P_0^ϵ) 問題を制約無しの最適化問題として再定義することができる．

$$(Q_1^\lambda): \qquad \min_{\boldsymbol{x}} \lambda\|\boldsymbol{x}\|_1 + \frac{1}{2}\|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}\|_2^2 \tag{6}$$

この (Q_1^λ) 問題は，統計的機械学習コミュニティでは Lasso として知られており，行列 $\mathbf{A}$ の各列が変化する特徴量を表し，複雑系システムの出力としてベクトル $\boldsymbol{b}$ が得られる場合，出力 $\boldsymbol{b}$ を表現する少数の特徴量の線形組み合わせを見つけることが目標である．尚，Lasso チームの Efron らにより 2004 年に提案された LARS (Least Angle Regression Stagewise) [12] は，(Q_1^λ) 問題の大域的最適解への解を保証するアルゴリズムである．

2.2.4 Pursuit 法の安定性

『Pursuit 法は近似的に (P_0^ϵ) 問題を解くことができるのであろうか?』という問いに対して，部分的ではあるが実験によって示された回答を紹介する．まず BPDN の安定性については，2006 年に Donoho らにより一定の結果 [13] が与えられ，$\boldsymbol{x}_0 \in \mathbb{R}^m$ が (P_1^ϵ) 問題の実行可能な解であり，スパース性制約 $\|\boldsymbol{x}\|_0 < (1 + 1/\mu(\mathbf{A}))/4$ を満たす場合，(P_1^ϵ) の解 $\boldsymbol{x}_1^\epsilon$ は $\|\boldsymbol{x}_1^\epsilon - \boldsymbol{x}_0\|_2^2 < \frac{4\epsilon^2}{1-\mu(\mathbf{A})(4\|\boldsymbol{x}_0\|_0 - 1)}$ に従うことが示された．一方，OMP 法よりシンプルな Thresholding 法についても安定性が示されている．(P_0^ϵ) 問題において，$\boldsymbol{x}_0 \in \mathbb{R}^m$ が $\|\boldsymbol{b} - \mathbf{A}\boldsymbol{x}_0\|_2 \leq \epsilon$ の実行可能な解であり，$\|\boldsymbol{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{A})}\frac{|x_{min}|}{|x_{max}|}\right) - \frac{\epsilon}{\mu(\mathbf{A})|x_{max}|}$ を満たす場合，$|x_{min}|$ 及び $|x_{max}|$ をサポート中の解 $\boldsymbol{x}_0$ の最小値及び最小値とすると，その解は，$\|\boldsymbol{x}_{THR} - \boldsymbol{x}_0\|_2^2 < \frac{\epsilon^2}{1-\mu(\mathbf{A})(\|\boldsymbol{x}_0\|_0 - 1)}$ に従う．ここで，スパース性への要求条件は，解 $\boldsymbol{x}$ の非零係数の最大値と最小値との比率，及びノイズレベル ϵ の両方に依存していることが分かり，BPDN のときとは異なる．ここで Thresholding 法のエラー限界は，BPDN の限界よりもより良いことが分かる．

尚，本分析は，**2.1.2**で述べた分析と同様に $\mathbf{A}$ の最悪特性に関するものであり，コヒーレンス，RIP，スパークともに，全て最悪値となっている．これに対し，確率的 RIP やコヒーレンスなどの，より緩和した手法を導入することが考えられ，近年研究により良い結果が示されている．例えば Ben-Haim らの研究成果 [14] を参照されたい．

3 ℓ_1 最小化最適化手法

前述の Pursuit 法による ℓ_0，ℓ_1 最適化手法は，高次元且つ大規模データの最適化には性能が良くない．そこで，ℓ_1 最小化に特化した最適化手法について紹介する．尚，ここでの対象問題は (Q_1^λ) 問題とする．この場合，構成する項は全て凸関数であることから大域的最小解を持つ．

3.1 反復縮退アルゴリズムベース法

反復縮退 (Iterative Shrinkage) アルゴリズムベース法は，要素毎の線形代数演算により実現され，それまでのアルゴリズムで必要な行列分解や線形最小二乗誤差を解く必要は無い．内点法等と比較すると反復回数は増加するが，各回の計算処理は大幅に削減されるという利点を有する．以下では，各手法で重要な役割を果たす proximal operator について先に紹介し，その後，代表的な手法について紹介する．

3.1.1 proximal operator

まず後述の手法で重要な役割を果たす proximal operator; $\mathrm{prox}_g(\boldsymbol{x}) \equiv \arg\min_{\boldsymbol{u}} \lambda g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2$ を定義する．ℓ_1 ノルムでは $g(\boldsymbol{x}) = \sum_{i=1}^{n} g_i(x_i) = \|\boldsymbol{x}\|_1$ なので $g(\boldsymbol{x})$ は分離可能であり，x_i で微分してその正負で場合分けすることで ℓ_1 最小化問題における proximal operator が導出される．

$$\mathrm{soft}(u, \lambda) \quad \equiv \quad \mathrm{sgn}(u) \max\{|u| - \lambda, 0\} \tag{7}$$

これは *soft-thresholding* または *shrinkage* operator と呼ばれる．入力値 u を理想の出力にマッピングするものであり， オリジナルの入力信号付近の値を零にし，閾値の外側の値は“ 縮退 ”する作用をもたらす．

3.1.2 座標降下法 (CD 法)

座標降下法 (Coordinate Descent :CD 法) は，(Q_1^λ) 問題を部分問題に分割して，x_i 要素以外の $\boldsymbol{x}$ の要素を全て固定して x_i を求める方式である．一般的な関数の場合は，微分可能でないとき，大域解に収束することは保証されない [15] [16] [17].

$$x_j \quad = \quad \mathrm{soft}\left(\sum_{i=1}^{n} a_{ij}\left(b_i - \sum_{p \neq j}^{m} a_{ip} x_p\right), \lambda\right) \tag{8}$$

単純に全座標を順番に処理する CD 法を Cyclic CD 法と呼ぶ [18]．収束までに各座標への処理が何度も行われるが，処理される座標の順序 (sweep pattern) が品質に大きな影響を与える．Wu らは 2008 年に全座標の中から最も小さな方向微分係数を有する座標を選択する sweep pattern の決定法を提案した [19]．一方，Li と Osher は，2009 年に Greedy CD 法を提案し [16]，(Q_1^λ) 問題を解く際に，エネルギーが最も減少する座標を適応的な sweep pattern 選択によりスパース性を実現する手法を提案した．これは，先の Wu らの手法と比較して座標選択における計算量が削減される利点がある．さらに Li らは，Basis Pursuit の解法について，Greedy CD 法と Bregman iterative method [20] を結合することで，正確性と効率性を両立した手法を提案した．Dhillon らも 2011 年に同様に Greedy CD 法を提案している [21]．以上述べた Greedy CD 法は，解がスパースな場合，ほとんど零の変数は処理途中で零のままとなり，問題次元が削減され，反復回数を削減することができる．

一方，Shalev-Shwarts らは，2009 年に Stochastic Coordinate Descent (SCD) 法を提案した [22]．この特徴は，各反復においてランダムに一つの座標 x_j を選択し更新する点にあり，実行時間の上界が問題サイズ (次元数 d) に比例する形で保証されるという点で特筆に値する．尚，大規模データの最適化に一般的に用いられる Stochastic Gradient Descent (SGD) 法 [23] は，ランダム選択されたサンプルに基づいて重みを更新する手法であるが，スパース解を与えることはできない．

3.1.3 近接点法 (ISTA, FISTA)

本節では，近接点法 (Proximal-Point Methods) を紹介する．(Q_1^λ) 問題において，$\boldsymbol{x}_{(k)}$ における 2 次近似を考えることで，目的関数は以下の制約無し BPDN 問題となる．

$$x_i^{(k+1)} \quad = \quad \mathrm{soft}\left(x_i^{(k)} - \frac{1}{\alpha^{(k)}}\nabla f(x^{(k)}), \frac{\lambda}{\alpha^{(k)}}\right) \tag{9}$$

この収束特性は $\alpha^{(k)}$ の決め方に依存する．例えば，Iterative Soft-Thresholding (Shrinkage) Thresholding (ISTA) [24] [25] では，リプシッツ連続勾配 L_f に連動した固定の $\alpha^{(k)}$ を選択する．この場合 $\mathcal{O}(1/k)$ 以下のほぼ直線的なレートで収束することを Beck らは示した [26]．さらにヘシアン $\nabla^2 f$ で模擬し $\alpha^{(k)}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \approx \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)})$ により最小二乗誤差が小さくなる $\alpha^{(k)}$ を求めることもでき，これは Barzilai-Borwein 方程式 [27] として良く知られている．また λ の選択方法は，固定値を使用する替わりに様々な戦略が提案された．Hale らは，ISTA の改善手法として fixed-point continuation method (FPC) を考案し [28]，各 λ_k に対して (Q_1^λ) 問題を解く替わりに，$\lambda = \lambda_0$ から開始して λ^* まで各反復時に $\lambda_{k+1} = \rho\lambda_k$ として徐々に減少させていく方式を提案した．Wright らも同様の手法を提案している [29] [24].

一方，ISTA の各反復は非常に簡易な計算であるが，実際には反復回数の観点から収束が遅いことが知られている．そこで Beck らは 2009 年に非常に良い収束レートを有する Fast ISTA (FISTA) を提案した [26]．同様に Newterov's Algorithm (NESTA) としても提案されている [30].

3.2 拡張ラグランジュベース法

本節では，反復縮退アルゴリズム法のもう一つの特別なクラスとして拡張ラグランジュ法 (Augmented Lagrange Multiplier: ALM)) を取り挙げ，(Q_1^λ) 問題に対する高速且つスケーラブルな手法を紹介する．

ALM は，繰り返し手法を用いて最適解とラグランジュ乗数を同時に推定する方法である [31]．ここで $h(\boldsymbol{x}) = \boldsymbol{b} - \mathbf{A}\boldsymbol{x}$ とし，ラグランジュ関数を $L_\mu(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{x}) + \frac{\mu}{2}\|h(\boldsymbol{x})\|_2^2 + \boldsymbol{y}^T h(\boldsymbol{x})$ と定義する．尚，$\mu > 0$ であり $\boldsymbol{y} \in \mathbb{R}^m$ はラグランジュ乗数である．乗数法 (Method of Multiplier) [32] [33] を使用することで，解を求める基本的な反復法は $\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} L_\mu(\boldsymbol{x}, \boldsymbol{y}_k)$, $\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \mu_k h(\boldsymbol{x}_{k+1})$ となる．μ_k は単調増加列であり，十分に大きい値のとき $\boldsymbol{x}^*$ と $\boldsymbol{y}^*$ は収束する．ここで第 1 ステップは，制約無し凸最適化問題であるため，元の制約付き最適化問題を直接解くことと比較して拡張ラグランジュ関数を最小化することが効率的に解くことができ，例えば **3.1.3**で述べた FISTA 等で効率的に解を導くことができる．以上述べた，ℓ_1 最小化問題を主問題として解く ALM を Primal ALM (PALM) と呼ぶ．ここで，$\mathbf{A}\boldsymbol{x} + \boldsymbol{e} = \boldsymbol{b}$ のように，観測信号 $\boldsymbol{b}$ に誤差が含まれる場合，つまり，(P_1^ϵ) 問題の場合，ラグランジュ関数は $L_\mu(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{y}) = \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 + \frac{\mu}{2}\|\boldsymbol{b} - \mathbf{A}\boldsymbol{x} - \boldsymbol{e}\|_2^2 + \boldsymbol{y}^T h(\boldsymbol{b} - \mathbf{A}\boldsymbol{x} - \boldsymbol{e})$, と書き換えられ，$\boldsymbol{x}$ と $\boldsymbol{e}$ を反復して交互に計算して求める．尚，$\boldsymbol{e}$ と $\boldsymbol{x}$ は FISTA 等で求めることができる．一方，$\boldsymbol{B}_1^\infty = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{B}\|_1^\infty\}$ と定義して，(P_1) 問題を双対問題 " $\max_{\boldsymbol{y}} \boldsymbol{b}^T\boldsymbol{y}$ subject to $\mathbf{A}^{\mathrm{T}}\boldsymbol{y} \in \boldsymbol{B}_1^\infty$ " と定義する Dual PLM (DPLM) も提案されている．拡張ラグランジュ関数は $\boldsymbol{x}$ をラグランジュ乗数として " $\min_{\boldsymbol{y},\boldsymbol{z}} -\boldsymbol{b}^T\boldsymbol{y} - \boldsymbol{x}^T(\boldsymbol{z} - \mathbf{A}^T\boldsymbol{y}) + \beta/2\|\boldsymbol{z} - \mathbf{A}^T\boldsymbol{y}\|_2^2$ subject to $\boldsymbol{z} \in \boldsymbol{B}_1^\infty$ " となり，交互法で解く．PALM と DALM の性能は，学習データ数と画像の次元数との関係に依存し，例えば一般的な顔画像認識では DALM が優れ，顔画像のアライメントが必要な場合には辞書の列数が少ないため PALM が優れている [31].

次に，先の ALM と類似する，Lee らにより提案された手法 [34] を紹介する．この手法はソースコードが公開されていることもあり，特に一般画像オブジェクト認識の多くの研究で使用されている．式 (3) に基底 $\boldsymbol{a}_i$ の二乗和が c 以下という条件を付与し，2 次式制約を持つ最小二乗誤差問題へと変形し，変数の交互法によりラグランジュ双対手法を用いて解を求める．具体

的には，$\mathbf{X}$ を固定して $\mathbf{A}$ を求める場合，ラグランジュ関数を $L(\mathbf{A}, \boldsymbol{\lambda}) = \mathrm{Tr}((\mathbf{B} - \mathbf{AX})^T(\mathbf{B} - \mathbf{AX})) + \sum_{i=1}^{m} \lambda_i(\sum_{j=1}^{n} \mathbf{A}_{j,i} - c)$ と定義する．ここで $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ とおくと，ラグランジュ双対は $\min_{\mathbf{A}} L(\mathbf{A}, \boldsymbol{\lambda}) = \mathrm{Tr}(\mathbf{B}^T\mathbf{B} - \mathbf{BX}^T(\mathbf{XX}^T + \boldsymbol{\Lambda})^{-1}(\mathbf{BX}^T)^T - c\boldsymbol{\Lambda})$ となり，共役勾配法等で最小化し $\mathbf{A}$ を導出する．これにより，双対問題を解くことで主問題より少ない数の変数の最適化に置き換えられ，例えば $\mathbf{A} \in \mathbb{R}^{1,000 \times 1,000}$ の場合，1,000 の双対変数の最適化で良い．

3.3 交互方向乗数法 (ADMM)

交互方向乗数法 (The alternating direction method of multipliers: ADMM) は，1975 年に Glowinski ら [35] 及び 1976 年に Gabay ら [36] により提案された．制約付き最適化の代表的な解法である乗数法 [32] は，分離可能な問題にそのまま適用しても元の問題の分離可能性を活かせない．そこで ADMM では，分離特性を有する Dual Ascent 法 (双対上昇法) と，優れた収束性を有する乗数法を混合した [37]．そのため，大規模最適化問題に対しても対応可能な分散処理に適した手法である．尚，ADMM は Douglas-Rachford splitting の他，数多くの最適化手法と密接に関係する．

$\boldsymbol{x} \in \mathbb{R}^n$，$\boldsymbol{z} \in \mathbb{R}^m$，$\mathbf{A} \in \mathbb{R}^{p \times n}$，$\mathbf{B} \in \mathbb{R}^{p \times m}$，$\boldsymbol{c} \in \mathbb{R}^p$ とし，且つ f と g が凸関数であることを仮定した上で，設定問題を“$\min_{\boldsymbol{x},\boldsymbol{z}}\ f(\boldsymbol{x}) + g(\boldsymbol{z})$ subject to $\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{z} = \boldsymbol{c}$”と定義する．等式制約問題“$\min f(\boldsymbol{x})$ subject to $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$”との差は，変数が $\boldsymbol{x}$ と $\boldsymbol{z}$ の 2 つに分割され，この分割に従った分割可能な目的関数を有することである．乗数法により拡張ラグランジュ関数を $L_p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \boldsymbol{y}^T(\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c}) + \frac{\rho}{2}\|\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c}\|_2^2$ で定義し，以下の反復を実行する．ここで以下にはより簡易的な表現である Scaled Form を示す．但し，$\rho > 0$ であり，$u = (1/\rho)y$ である．

$$
\begin{cases}
\boldsymbol{x}^{(k+1)} \equiv \arg\min_{\boldsymbol{x}}\ f(\boldsymbol{x}) + \frac{\rho}{2}\|\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{z}^{(k)} - \boldsymbol{c} + \boldsymbol{u}^{(k)}\|_2^2 \\
\boldsymbol{z}^{(k+1)} \equiv \arg\min_{\boldsymbol{z}}\ g(\boldsymbol{z}) \\
\qquad\qquad + \frac{\rho}{2}\|\mathbf{A}\boldsymbol{x}^{(k+1)} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c} + \boldsymbol{u}^{(k)}\|_2^2 \\
\boldsymbol{u}^{(k+1)} \equiv \boldsymbol{u}^{(k)} + \mathbf{A}\boldsymbol{x}^{(k+1)} + \mathbf{B}\boldsymbol{z}^{(k+1)} - \boldsymbol{c}.
\end{cases}
\tag{10}
$$

一方，式 (10) に対する標準的な乗数法は $(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1}) \equiv \arg\min_{\boldsymbol{x}} L_\rho(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}_k), \boldsymbol{u}_{k+1} \equiv \boldsymbol{u}_k + \rho(\mathbf{A}\boldsymbol{x}_{k+1} + \mathbf{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$ であった．拡張ラグランジュでは，2 つの主問題変数の観点を同時考慮して最小化を行うのに対して，ADMM では $\boldsymbol{x}$ と $\boldsymbol{z}$ を交互に更新する．これが交互方向と呼ばれる所以である．尚，収束性については [37] 等を参照されたい．

3.4 大規模高次元データへの対応

アプリケーションがより大量の高次元データを活用するにつれて，スケーラブルな最適化手法への期待はより高まっている．一方で，プロセッサの速度向上は止まりつつあり，その替わりにコア数の増加が進んでいる．これまで，ℓ_1 最適化に対して逐次処理による最適化手法は多数検討されているが，並列化に対応したアルゴリズムはまだ多くは存在しない．以下では，大規模高次元データへの対応を見越した最適化手法についていくつか紹介する．

3.4.1 分散処理の基本的考え方

$\boldsymbol{x} \in \mathbb{R}^m$，$\mathbf{A} \in \mathbb{R}^{n \times m}$ とし，問題を“$\min_{\boldsymbol{x}}\ \mathcal{F}(\boldsymbol{x}) = f(\mathbf{A}\boldsymbol{x}, \boldsymbol{b}) + \lambda g(\boldsymbol{x})$”と定義する．ここで $f(\mathbf{A}\boldsymbol{x}, \boldsymbol{b})$ は滑らかな損失関数とし，$g(\boldsymbol{x})$ は **1.4**で述べた何らかの構造を有する正則化項とする．ここで $\boldsymbol{x}_s$ が $\boldsymbol{x}$ の s 番目ブロックと定義すると，もし $f(\mathbf{A}\boldsymbol{x}, \boldsymbol{b}) = \sum_{s=1}^{S} f_s(\boldsymbol{x}_s)$ であるならば $f(\mathbf{A}\boldsymbol{x}, \boldsymbol{b})$ は（ブロック）分離である．一方で，正則化項 $g(\boldsymbol{x})$ として一般的に使用される ℓ_1 ノ

ルムや $\ell_{1,2}$ ノルム，Huber 関数や Elastic Net 関数は，いずれも $g(\boldsymbol{x}) = \sum_{b=1}^{B} g(\boldsymbol{x}_b)$ を満たす．これに基づき，行列 $\mathbf{A}$ を分割後分散し，また観測ベクトル $\boldsymbol{b}$ または求めるべき解 $\boldsymbol{x}$ を分散的に配置し，$\min_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x})$ 問題に対して $\mathbf{A}\boldsymbol{x}$ または $\mathbf{A}^T\boldsymbol{b}$ を分散的に解くことで最適解を求める．

ここで行列分割について説明すると，“行分割”の場合は $\mathbf{A}$ を部分行列 $\mathbf{A}_{(1)}, \cdots, \mathbf{A}_{(n)}$ から構成し，“列分割”の場合は $\mathbf{A} = [\mathbf{A}_1\mathbf{A}_2, \cdots, \mathbf{A}_m]$ とする．但し，$\mathbf{M}_{(i)}$ のように表記する場合は行-行列である．行分割は，比較的次元低い特徴ベクトル $\boldsymbol{x}$ を有しながら，観測信号（サンプル) 数が非常に大きい場合に適している．一方で，列分割は，比較的少ないデータサンプルに対して，それらが極めて高次元な特徴ベクトルにより構成されている場合に適している．

実際の計算においては，中間ノードと複数の計算ノードから成るシステムを構築する．そして，中間ノードは，各ノードに向けて行列や観測情報，計算結果を分配 (*broadcast/scatter*) 後，各計算ノードの計算 (*compute*) 結果を収集 (*gather*) し，集約計算 (*reduce*) を行い，再度分配するという処理を繰り返す．例えば行分割の場合は，中間ノードは i 番計算ノードの計算結果 $\mathbf{A}_{(i)}^T\mathbf{A}_{(i)}\boldsymbol{x}^{(k)}$ を *gather* し，$\sum_{i=1}^{N} \mathbf{A}_{(i)}^T\mathbf{A}_{(i)}\boldsymbol{x}$ を用いて *reduce* し，$k+1$ 回目の $\boldsymbol{x}^{(k+1)}$ を導出することで，再度 $\boldsymbol{x}^{(k+1)}$ を *broadcast* する．

3.4.2 並列近接点法 (Parallel FISTA)

正則化項と損失項がブロック分離可能なとき，**3.1.3**で述べた ISTA や FPC，FISTA アルゴリズムなどのアルゴリズムを並列化することができる．Peng らは 2013 年，前述の FISTA を並列化する手法を提案した [18]．例えば列分割の場合は，各ノード j $(0 < j \leq M)$ は，$\mathbf{A}_j$ と全体の $\boldsymbol{b}$，最新の $\boldsymbol{x}^{(k)}$ の一部 $\boldsymbol{x}_j^{(k)}$ を保持し，ループ処理に入る前に後続の Soft 関数処理で使用する $\delta\mathbf{A}_j^T\boldsymbol{b}$ を一度だけ計算し保持しておく．そして計算ループでは，各ノードは $\mathbf{A}_j\boldsymbol{b}_j^{(k)}$ の演算を行い，それを収集した中間ノードは $\boldsymbol{y} = \sum_j^M \mathbf{A}_j\boldsymbol{b}_j^{(k)}$ 計算後，それを分配し，各ノードは式 (9) に対応する $\boldsymbol{x}_j^{(k+1)} = \mathrm{soft}(\boldsymbol{x}_j^{(k)} - \delta\mathbf{A}_j^T\boldsymbol{y} + \delta\mathbf{A}_j^T\boldsymbol{b}, \lambda\delta)$ による Soft 関数処理を行い，以後，同様の繰り返し処理を続ける．尚，行分割の場合も同様に考えることができる．

3.4.3 並列・分散座標降下法 (Parallel/Distributed CD 法)

3.1.2で説明した座標降下 (CD) 法は，反復時に 1 つの座標を更新するため，簡易且つ効果的なアルゴリズムである．また SGD 法のようにパラメータを調整する必要も無いことが利点である．一方，近年のデータの大規模化に対して SGD 法に対する Parallel SGD [38] 等が提案された．これらはサンプルデータについて並列化を行なうであり，今後のビッグデータに対応する手法となる．このような中，CD 法についても，Yuan と Lin は大規模スケールの ℓ_1 正則化問題に対して CD 法を含むいくつかの方式に対するシミュレーション実験比較を行っているが [39]，これは逐次処理アルゴリズムのみを対象としている．それに対し Shwarts と Tewari により，理論的且つ十分な実験結果から CD 法の高次元データに対して非常に良い性能を示すことが明らかにされている [22]．そこで，Bradley らは，2011 年に SCD 法の並列手法として Parallel Stochastic Coordinate Descent 法 (Parallel SCD) を提案した [40]．ここでは，各反復において，可能な組み合わせからランダムに P 個の x_{i_j} のサブセットを選択する．そして，各プロセッサで計算した x_{i_j} に対応する更新分 δx_{i_j} を集め，$\sum_{i_j \in \mathcal{P}} \delta x_{i_j}$ を $\boldsymbol{x}$ の更新分 $(\Delta\boldsymbol{x})_j$ とする．さらに，Scherrer らは 2012 年に GenCD 法を提案し，Parallel Coordinate Descent (PCD) 法に対する一般化を行なった．GenCD 法の特殊ケースとして，Li らが提案した Greedy CD 法，前述の Parallel SCD 法，等が含まれる．一方，Richtárik と Martin Takáč は，2011 年，GPU アクセラレーションによる Greedy ランダム CD 法の並列化手法について提案した [41]．また，Scherrer らは，2012 年に Parallel Block-Greedy Coordinate Descent (Parallel Block-Greedy CD) 法を提案した [42]．入力特徴ベクトルを B 個に分割し，各反復では，この中かからランダムに P 個のブロックを選

択し，各ブロックから一つの特徴を選択した．ここでの選択は，目的関数を減少させる度合いを推定したて行われる．

さらに，大規模データの分散処理を強く意識したスケーラブルな分散 CD 法として，2013 年に Rihtarik らが Hydra を提案し [43]，3TB のサイズを有する行列 $\mathbf{A}$ に対する実験を行い，その有効性を示している．提案では，複数ノード間をリング状に結合しデータ共有を行なう通信プロトコルを提案し，その有効性も示している．さらに，Peng らも，反復回数の少ない Greedy CD 法のアルゴリズム GRock を 2013 年に提案した [18] ．また，この他，Elad らにより提案された並列化手法 [44] 等もある．

3.4.4 分散交互方向乗数法 (Distributed ADMM)

3.3で述べた ADMM の分散処理への拡張が検討されており [37]，例えば，式 (10) の第 1 式の $\boldsymbol{x}$ 更新式については $\boldsymbol{x}_i^{(k+1)} \equiv \arg\min_{\boldsymbol{x}} f_i(\boldsymbol{x}_i) + \frac{\rho}{2}\|\mathbf{A}_{(i)}\boldsymbol{x} + \mathbf{B}_i\boldsymbol{z}_i^{(k)} - \boldsymbol{c}_i + \boldsymbol{u}_i^{(k)}\|_2^2$ となる．このような中，Mateos らは (Q_1^{λ}) 問題 (Lasso) に対する分散手法を検討した [45]．分割数を N とした行分割の場合，$\boldsymbol{x} \in \mathbb{R}^n$，$\boldsymbol{z} \in \mathbb{R}^n$，$\mathbf{A}_{(i)} \in \mathbb{R}^{n_i \times m}$，$\mathbf{b}_i \in \mathbb{R}^{n_i}$，$\sum_i n_i^N = n$ とすると，設定問題は，“$\min_{\boldsymbol{x},\boldsymbol{z}}\ (1/2)\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{z}\|_1$ subject to $\boldsymbol{x} - \boldsymbol{z} = \mathbf{0}$”となり，最終的に式 (10) は以下となる．

$$\begin{cases} \boldsymbol{x}_i^{(k+1)} \equiv \arg\min_{\boldsymbol{x}} (1/2)\|\mathbf{A}_{(i)}\boldsymbol{x} - \boldsymbol{b}_i\|_2^2 \\ \qquad\qquad\qquad\qquad + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{z}^{(k)} + \boldsymbol{u}_i^{(k)}\|_2^2 \\ \boldsymbol{z}^{(k+1)} \equiv \mathrm{soft}_{\lambda/\rho N}(\bar{\boldsymbol{x}}^{(k+1)} + \bar{\boldsymbol{u}}^{(k)}) \\ \boldsymbol{u}_i^{(k+1)} \equiv \boldsymbol{u}_i^{(k)} + \boldsymbol{x}_i^{(k+1)} - \boldsymbol{z}^{(k+1)}. \end{cases} \tag{11}$$

尚，$\bar{\boldsymbol{a}}$ は平均値を示す．$\boldsymbol{x}_i^{(k+1)}$ の導出は Tikhonov 正則化最小誤差 (リッジ回帰) 問題となり，解析的に $\boldsymbol{x}^{(k+1)} \equiv (\mathbf{A}_{(i)}^T\mathbf{A}_{(i)} + \rho\mathbf{I})^{-1}(\mathbf{A}_{(i)}^T\boldsymbol{b}_i + \rho(\boldsymbol{z}^{(k)} - \boldsymbol{u}_i^{(k)}))$ と求まる [37].

4 辞書学習

辞書の構成法には，大きく解析ベースと学習ベースアプローチがある．前者は数学的モデルに基づき解析的な構成によりモデルを表現するもので Wavelet，Curvelets 等がある．一方，後者は，学習データから辞書を推定する機械学習技術に基づき，例えば PCA などが代表例である．この辞書の作成にはコストがかかり，処理対象信号の次元や辞書のサイズに制限が加わるが，対象とする信号に対してより極め細やかな辞書を作成することができる利点がある．本稿では，学習ベース辞書構築手法について紹介する．まず，学習用データを $\boldsymbol{b}_i (i = 1, \cdots, M)$ が，ある未知のモデル $\mathcal{M}_{\{\mathbf{A},k_0,\alpha,\epsilon\}}$ から生成されたものとする．このとき最適な辞書 $\mathbf{A}$ を求める問題“$\min_{\mathbf{A},\{\boldsymbol{x}_i\}_{i=1}^M} \sum_{i=1}^M \|\boldsymbol{x}_i\|_0$ subject to $\|\boldsymbol{b}_i - \mathbf{A}\boldsymbol{x}_i\|_2$”を考える．ここで，$\boldsymbol{b}_i$ について $\mathbf{A}$ を用いてスパースな係数 $\boldsymbol{x}_i$ で表現することを考え，同時に辞書を見つけることを考える．さらに，もし k_0 個またはそれ以下の非零係数で表現できればモデル $\mathcal{M}$ を見つけることができるものと考える．ここで，スパース性を制約条件として，辞書学習問題は以下の最適化問題を解くことに帰着される．

$$\min_{\mathbf{A},\{\boldsymbol{x}_i\}_{i=1}^M} \sum_{i=1}^M \|\boldsymbol{b}_i - \mathbf{A}\boldsymbol{x}_i\|_2^2 \ \text{ subject to } \|\boldsymbol{x}_i\|_0 \le k_0 \tag{12}$$

4.1 代表的な辞書学習法

多数の取り組みの中から，いくつかの代表的な成果について紹介する．Engan らは，1999 年に MOD (Methods of Optimal Direction) を発表した [46]．式 (12) について，辞書 $\mathbf{A}$ と誤差を固定の上で，$\boldsymbol{x}_i$ のうち，非零係数の数を最小化する問題と見なし，交互最小化問題として解く．具体的には，k 番目のステップにおいては，$k-1$ 番目ステップで得られた辞書 $\mathbf{A}^{(k-1)}$ を用いて，M 個の (P_0^ϵ) 問題を解くことを考える．各 $\boldsymbol{x}_i$ については，$\mathbf{A}^{(k-1)}$ と各 $\boldsymbol{b}_i$ を用いて式 (12) の P_0^ϵ 問題により“$\mathbf{A}^{(k)} = \arg\min_{\mathbf{A}} \|\boldsymbol{b}_i - \mathbf{A}\boldsymbol{x}_i\|_F^2 = \mathbf{B}\mathbf{X}^{(k)^T}(\mathbf{X}^{(k)}\mathbf{X}^{(k)^T})^{-1} = \mathbf{B}\mathbf{X}^{(k)^+}$”として $\mathbf{X}^{(k)}$ を得る．

Aharon らにより 2006 年に発表された K-SVD [47] は，その後の辞書学習法や SC 応用分野など，多数の研究に大きな影響を与えている．初期辞書 $\mathbf{A}_0$ と学習データを入力として反復処理を行ない，学習データに基づいて基底 $\boldsymbol{a}_i$ を学習する．具体的には，1) 入力信号 $\boldsymbol{b}_i$ に対する Pursuit 法を用いたスパース係数の導出処理と，2) スパース係数が与えられた上での一つの基底 $\boldsymbol{a}_i$ の更新処理，を交互に繰り返す．基底の更新においては，係数のスパース制約を維持するために，対象基底 $\boldsymbol{a}_i$ を用いる $\boldsymbol{x}_i$ と $\boldsymbol{b}_i$ のみを用いて，式 (12) 第 1 項の損失項の最小化を行なう点がポイントである．最小化にあたっては，残差誤差に対して SVD を施すことにより得られる左右特異値ベクトルを用いて，基底 $\boldsymbol{a}_i$ と対応するスパース係数 $\boldsymbol{x}_i$ を更新する．注意すべき点は，MOD と K-SVD ともに，損失項の大域的最小化を保証できない点である．またデータや辞書サイズの次元の増大により処理量が急激に増加するため，低次元への制限がある点，また，シフト/回転/スケール等の変化に対する耐性が低い点が挙げられる．

さらに Rubinstein らは，解析的辞書と学習辞書の中間に位置する辞書の構築法として，解析的辞書 $\mathbf{\Phi}$ とスパース性を有する学習辞書 $\mathbf{A}_s$ とから構成される $\mathbf{A} = \mathbf{\Phi}\mathbf{A}_s$ を用いた $\mathbf{A}_s$ の辞書学習法としてスパース K-SVD (Sparse K-SVD) を提案した [48]．これにより，K-SVD では考慮されなかった辞書のスパース化も実現でき，表現能力だけでなく辞書構築の効率化も達成できる．

一方，再構成する表現性能ではなく，識別性能を高める学習方法 (Discriminative Training) が提案されている．Mairal らは，スパースな再構成とともに線形予測モデルを最適化するスパース辞書学習モデル (Supervised Dictionary Learning) を提案した [49]．また Bradley らは，従来の ℓ_1 ノルムではなく微分可能なスパース事前確率を提案し，学習誤差を最小化するための辞書学習方法を提案した [50].

4.2 大規模高速辞書学習法

前述の K-SVD をはじめとする辞書学習法は，各反復で全ての学習データにアクセスするバッチ処理を基本としている．そのため大規模な学習データを処理することは難しく，時々刻々変化する動的データへの対応も難しい．このような問題に対して，Miral らは 1 回の処理で 1 つの要素にしかアクセスしないオンライン手法を提案した [51]．学習時に全データが必要無いことから大規模データにスケール可能であり，収束性も示されている．この成果はそれ以後の研究に影響を与え，例えばグループ ℓ_1 ノルムをベースとしたオンライン学習法等も提案された [52].

一方，Xiang らは，大規模辞書を用いた高次元データのスパース係数導出及び辞書学習法について提案している [53]．具体的には，1) スパース係数を求めるべきデータと基底との関係性に基づき，ℓ_1 最適化に用いる不要な基底を適応的に省略し，2) また階層化辞書モデルを導入し，各階層で使用するデータは，圧縮センシングの原理に従い，Random Projection により次元を削減して各階層で学習を行なう．これにより小さな問題に分割することで，大規模な学習処理を可能とした．さらに Sindhwani らは，実際の大規模データを対象とした検討として，大量のドキュメント情報を対象とした解析において，非負スパースコーディングとスパース辞書学習法の両方を同時に行なう実装を行なった [54]．ここでは，非負 OMP 法と非負 Lasso を Hadoop

により並列化し，辞書学習法は K-SVD と類似のスパース制約を持つ学習法により実現した．これにより，汎用のクラスターを用いて 1 億以上の行を有し 10 億の非零要素を持つ行列を数時間以内に最適化できることを確認している．

4.3 行列分解（Matrix Factorizations）との関連

SC 及び辞書学習は $\mathbf{B} \approx \mathbf{AX}$ の行列分解と等価である．行列分解は，主成分分析 PCA，独立成分分析 ICA，非負値行列因子分解 NMF(Non-negative matrix factorization) などが代表的である．特に，NMF は構成行列がスパースになることからスパースコーディングとの関連性は高く，また近年，信号処理やテキスト分類など，負の成分を持たない信号に対する応用が研究されている．しかしながら，NMF には明確なスパース制約はなく，非負に起因して結果的にスパースになり易いという性質を有する．これに対して Hoyer は ℓ_1 ノルムと ℓ_2 ノルムの比率に基づくスパース制約付き NMF を提案 [55] し，Peharz らは ℓ_0 ノルム制約付き NMF を提案している [56]．一方，K-SVD 考案者の Aharon は，K-SVD に非負制約を付与した Non-negative-KSVD を提案している [57] が，K-SVD ほど性能が良くないことが報告されている．NMF の大規模データ対応については Scalable CD 法ベースの並列行列分解手法 [58] 等多数の提案されている [59] [60]．NMF と SC との関係は行列導出方法に違いがあるものの，その効果は類似しており，今後も相互に影響しながら発展していくものと考えられる．

5 スパースコーディングの応用

スパースコーディングは広範な分野に適用されその効果を示しているが，本節では，その中のいくつかの応用分野にフォーカスを当て，代表的な研究成果について紹介する．

5.1 画像ノイズ除去，画像超解像，他

画像ノイズ除去 (Denoising) への適用は，(P_0^ϵ) 問題を満たす解 がノイズ除去された理想的なクリーンな画像を表現するスパース表現であるという仮説に基づく．画像修復の数多くの研究の代表として，ローカルパッチにおけるスパース性と冗長性の考慮と，全体領域での平滑化とベイズ的な最適化によりノイズ除去を行なう Elad らの提案 [61] が挙げられる．Pursuit 法と K-SVD との統合によりノイズ除去を行なうもので，同著者らによるカラー画像へ拡張 [62] やマルチスケール辞書による改善やビデオ修復への拡張 [63]，パッチ間の類似性を考慮したグループ ℓ_1 ノルム制約に基づく LSSC [64] を含め，その後の K-LLD [65] や他多数の研究に影響を与えた．

画像超解像 (Super-resolution) に対する最初の代表的な成果は，低解像度画像と高解像度画像は同一のスパースコードを有するという仮説に基づき，まず低解像度用辞書を学習し，これを用いて得られる処理対象画像のスパース係数に対応する高解像度画像を用いて超解像度画像を生成する Yang らの提案である [66]．さらに同著者らは，前述の LSSC で考慮したパッチ間の類似性を複数スケールにも拡張し，グループ ℓ_1 制約に基づいて超解像度画像を生成する手法を提案した [67]．その後，高周波数成分の学習方法の改善により品質向上を実現する Zeyde らの手法 [68] や，異なる解像度画像の組とその間のマッピング関数を同時に学習する Wang らの手法 [69]，辞書を部分辞書に分割する Dong らの手法 [70] 等，数多くの成果が発表されている．

その他，画像修復 (Inpainting) や画像分離 (Separation)，画像圧縮 (Compression) に適用され，最高水準の成果を上げている．これらの詳細については [4] に詳しい．

5.2 顔画像分類

SC を用いて識別特徴量を学習する Sparse Representation Classification (SRC) の研究が多数進められており，特に顔画像認識への応用で成功している．Wright らは学習用画像セットを SC 用辞書とし，認識問題を認識対象画像のスパース係数から求める問題に帰着させ [71]，その後の多数の研究に影響を与えた．まず，クラス数を K とし，$\boldsymbol{x}_{kj} \in \mathbb{R}^N$ は k 番目クラスの j 番目学習画像からの特徴ベクトルを示すものとする．k 番クラスからの特徴ベクトル行列 $\mathbf{A}_k \in \mathbb{R}^{N \times n_k}$ を $\mathbf{A}_k = [\boldsymbol{x}_{k1}, \cdots, \boldsymbol{x}_{kn_k}]$ とし，$\mathbf{A} \in \mathbb{R}^{N \times \sum_{k=1}^{K} n_k}$ を $\mathbf{A} = [\mathbf{A}_1, \cdots, \mathbf{A}_K]$ とする．ここでテスト画像 $\boldsymbol{b} \in \mathbb{R}^N$ を，係数 $\alpha_{kj} \in \mathbb{R}$ を用いて $\boldsymbol{b} = \sum_{k=1}^{K} \sum_{j=1}^{n_K} \alpha_{kj} \boldsymbol{x}_{kj} = \mathbf{A}\boldsymbol{\alpha}$ のように学習ベクトルの線形和で表現する．尚，$\boldsymbol{\alpha}$ は $\boldsymbol{\alpha} = [\alpha_{11}, \cdots, \alpha_{2n_1} | \ \cdots \ | \ \alpha_{K1}, \cdots, \alpha_{Kn_K}]^T$ である．ここでの仮説は，十分な数の $\mathbf{A}_k$ がある場合，k_0 番目クラスに属するテスト画像 $\boldsymbol{b}$ は，$\mathbf{A}_{k_0}$ により張られる線形空間内に近似的に位置し，クラス k_0 に属さない係数 $\boldsymbol{\alpha}$ のほとんどが零になることである．ここで $\hat{\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}} \|\boldsymbol{x}\|_1$ subject to $\boldsymbol{b} = \mathbf{A}\boldsymbol{\alpha}$ を解くことで $\hat{\boldsymbol{\alpha}}$ を導出し，$\hat{\boldsymbol{\alpha}}$ のうちクラス k 以外に対応する係数を零にして得られる係数 $\Pi_k(\hat{\boldsymbol{\alpha}})$ からの再構成画像の残差誤差を $\boldsymbol{r}_k(\boldsymbol{b}) = \|\boldsymbol{b} - \mathbf{A}\Pi_k(\hat{\boldsymbol{\alpha}})\|_2$ で計算し $\arg\min_k r_k(\boldsymbol{b})$ によりクラス識別を行なう．尚，$\Pi_k : \mathbb{R}^n \to \mathbb{R}^n$ は，k 番目クラスに対応する係数のみを抽出する演算子である．

拡張研究として，Elhamifar らは辞書に含まれる学習データにはブロック構造があることに着目し，ℓ_1 または ℓ_2 ノルムの評価にブロック構造を考慮した方法を提案している [72]．さらに，複数の異なるバイオメトリック特徴を用いた識別問題を対象として，マルチモーダル多変数スパースコーディングによるクラス識別手法も提案されている [73] [74]．複数のモダリティを考慮し，同一クラスの異なるモダリティに対するスパース性を考慮していることから多変量 Lasso とも呼ばれる．さらにノイズが含まれる場合も考慮し，特にオクルージョン項のスパース性については Wright [71] らや Candes ら [75] の検討結果に基づいている．本問題を解くには ADMM を使用している．

一方，ℓ_1 最小化の計算量を考慮すると ℓ_1 スパース制約が解の正則化において必須であるかどうかという指摘 [76] に対して，Zhang らは，必ずしも ℓ_1 スパース制約は必要ではなく ℓ_2 ノルム制約のほうがよい性能を示すことを示している [77]．同時に，学習データ数が少ない場合には，識別性能において他のクラスの基底を用いた協調表現の重要性についても示している．これは，SRC では $\mathbf{A}$ が過完備の場合を想定しているものの，学習データ数は一般に顔画像などの $\boldsymbol{x}$ の次元数より少ないことに起因している．そこで，圧縮センシング (Compressive Sensing) により次元削減する手法が提案されている [78]．近年，Random Projection が高い性能を示すことが示されており [79]，どの特徴量を使用するかはそれほど重要ではないことが報告されている．

5.3 一般オブジェクト画像分類

前節の SRC は，画像オブジェクトに対して直接作用するため，位置合わせした顔や数字などの単純な信号に限定され，物体やシーン識別などの一般画像についてはうまくいかない．そこで，信号のスパース係数を SVM 等の一般識別器で利用することが検討されている．まず最初に SC が大きな成果を挙げたのは，2009 年に Yang らにより提案された ScSPM である [80]．局所特徴の集合を示す Bag-of-Features (BoF) モデルと，特徴の空間レイアウトに関する情報を記述するため周期的に画像を分割する空間ピラミッドマッチング (Spatial pyramid matching: SPM) カーネルを用いている．従来の SPM カーネルを用いた非線形 SVM 識別法のベクトル量子化（VQ）を SC に一般化するとともに，特徴ベクトル平均値を使用した Pooling 法を，最大値を採用する Max Pooling 法に変更した．スパースコード統計情報に基づき線形 SPM カーネルを使用することで，識別精度を向上しながら学習時の処理量及びメモリ量を大幅に削減した．これに対して，Gao らは，ScSPM では特徴量間の相互依存関係を無視するため，類似の

局所特徴量でさえ異なるスパースコードが導出されることがあることに着目し，量子化誤差を削減し，局所特徴量間のスパースコードの一致性も維持することが可能な LScSPM を提案した [81]．一方，Yu らは，高次元データでもしばしば非常に小さな次元上に位置することに着目し，局所条件を加味したマッピング関数とアンカーポイントの組合せ符号化（Local Coordinate Code: LCC）に基づいて，多様体上の非線形関数を線形関数で近似する方法を提案した [82]．Wang らは，LCC や LScSPM と同様に局所条件に着目し，結果としてスパース性を実現する手法 Locality-constrained Linear Coding (LLC) を提案した [83]．局所性に着目することで，パッチ間の関係性を維持し，局所領域に位置する近い画像に対しては，コードブックから類似した基底が抽出されるようにした．さらに，Yang らは，ScSPM は高い性能を示すが，SC のための学習およびテストの処理量が高く，特に過完備辞書の学習には高い処理量を要する点に着目し，小さな複数の辞書の混合モデルを用いて，SC の処理量を効率的に削減する混合モデル手法を提案した [84]．また同著者らは，Back-Projection 法により空間ピラミッド内のスパースコードを Max Pooling から抽出することで，画像レベルでの特徴に関する識別誤差を最小化する教師あり辞書学習法について提案した [85]．特に，異なるスケールのスパースコードを畳込んだ階層モデルを構築し，複数の空間スケールにまたがる Pooling 処理により，空間特性と位置移動性を両立した特性を得ることができる．

一方，Yu らは，2 階層モデルを採用し，局所領域内のパッチ間の高次依存関係をモデル化した [86]．第 1 層は各パッチで符号化し，第 2 層は同じグループ（同じ領域）にあるパッチセットを同時に符号化する．各パッチ用と低レベルコードワード間のパターン依存関係を表現するパッチセット用の 2 つのコードブックを持ち同時に学習する．さらに He らは，多階層モデルとして Deep Sparse Coding (Deep SC) を提案した [87]．具体的には，複数階層間を”sparse-to-dense” モジュールにより結合させ，局所空間 Pooling と隣接パッチ画像間の関係を維持した次元圧縮により空間スムーズ性を考慮した情報を埋め込む処理を実現している．複数階層からの空間ピラミッド Pooling 特徴を用いて識別を行なうことで性能向上を実現している．

最後に，大規模データに対して，Lin らは大規模画像データセット (1000 クラス 1.2 million 画像) である ImageNet を対象として高速化を実現している [88]．具体的には，HOG 等の特徴量抽出と前述の LLC 及び SVC による符号化，さらには SPM によるプーリング処理を行なう場合，最大 208 日必要と見積もられる処理を，Hadoop で 120 ワーカーを使用することで最大 2 日程度まで時間短縮を実現した．一方，SVM の学習では 1000 クラスに対して 1 対 1000 のバイナリー識別を並列化することで，8 コア 12 台の計算機を使用して 250 日必要と見積もられた処理を，1 週間以内で実現した．尚，ここでは，広く使用される LibSVM などのライブラリでは実行できないため，Averaging Stochastic Gradient Descent (ASGD) 法を用いている．

5.4 オーディオ信号及び楽曲処理

DFT や DCT，あるいは DWT などの可逆な変換を用いず，SC を用いたオーディオ信号処理方法が提案されている [89]．オーディオ信号のノイズ除去技術においては，楽曲の各パートはスパース表現により良く表現されるのに対して，ノイズはスパース表現ではうまく表現できないことを利用して，信号をスパース表現に変換し，より小さな係数を削除した上で再構成することで，重要な部分だけを取り出すことが可能となる．ここで，楽曲に関する構造上の事前知識を導入し，ベイジアンフレームワーク上で実現することで，共鳴構造を垂直方向の周波数構造で表現し，音調については水平方向の構造で表現することで，残のノイズの分散とともにモデル化することが可能となる [90]．

一方，ポリフォニー（多声）音楽の採譜に対して時間ベース及び周波数ベースの手法として SC の使用が提案されたのは，2001 年の Abdallah らの提案が最初である [91]．これは同著者らの研究 [92] やシフト不変 SC により時間軸上での信号を直接的に扱う Plumbley らによる手法 [93] により改善され，また多数の研究に影響を与えた [94]．さらに，各係数に意味を持つ

オーディオ信号についても検討されている．例えば，キーボード楽器等，複数の候補音符の中から，一度に数個のみしか構成されない場合である．この場合，各音符は時間—信号スペクトル上でスパースな信号を形成するため，K-SVD などを使用してデータから基底を学習することで，より効率的な表現が可能となる [92]．但し，各組合せ対する大規模なデータベースが必要り，また単一音符が学習データに現れなかった場合に各音符を表現する辞書ベクトルが無いという問題がある．そこで Genussov らは，88 音府の全ての各音符から構成される初期辞書を作成し，K-SVD を発展させた辞書学習フェーズでは，周波数成分を変更することなくパワー成分だけ学習する Musically-structured 辞書学習法を提案し，多重基本周波数の推定方法を提案した [95]．この他，Lee らによるピアノ楽曲の複数音程推定の提案 [96] や，グループスパース性と非負 SC による楽曲自動採譜の提案 [97] がある．さらに，オーディオ分野への応用としては，音の時間変調を利用した SRC ベースの楽曲ジャンル分類技術 [98] や，ノイズロバストな自動スピーチ認識 [99] ，サウンド分離 [100] [101] 等の多数の研究が進められている．

5.5 変化・イベント検知

スパースコードをイベント検知に適用する検討が多数行なわれている．共通する考え方は，与えられた辞書 $\mathbf{A}$ を用いて，スパース制約の下で観測ベクトル $\boldsymbol{b}$ の係数 $\boldsymbol{x}$ を算出する時，$\boldsymbol{x}$ が十分にスパースな場合は普通のイベントとし，スパースでなく多数の係数が出現する場合には，辞書 $\mathbf{A}$ で表現できない異常なイベントとして検出するというアプローチである．Cong らは，これに加え，辞書の選択手法や辞書の自動更新手法について提案し，監視カメラ映像を用いてその有効性を評価している [102]．Zhao らも，映像を時空間の 3D 直方体としてとらえ，時間，空間両軸で観測窓をスライドさせながら SC を行なう手法，及び辞書の自動更新を技術を含む同様の提案をしている [103]．一方，Adler らは，データにノイズが混在する場合を考慮して，グループ ℓ_1 ノルムを用いた最適化による異常検知法を提案して [104] おり，心電図を対象とした検証を行なっている．最適化には ADMM を用いている．最後に大量のドキュメント情報に対する新しいトピックを有する新規ドキュメントの検出手法についても紹介しておく．基本的な考え方は同じであるが，スパースコーディングフェーズと辞書学習フェーズをそれぞれ並列化し，分散 ADMM により分散的に処理する手法を Kasiviswanathan らは提案している [105]．128 及び 256 のプロセッサークラスターを用いて，約 41 日間の Tweet を対象として実験を行い，処理量及び通信量オーバヘッドの評価を行なっている．

6 まとめ

スパースコーディングの基礎からその応用まで，紙面が許す限り紹介した．内容が広範なため，全てを紹介することは不可能であるが，本稿がスパースコーディングに関心のある研究者にとって少しでも役立てば幸いである．

References

[1] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[2] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[3] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, December 1998.

[4] Michael Elad. *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[5] Emmanuel Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23:969–985, 2007.

[6] D.L. Donoho and M. Elad. Optimally sparse representations in general (non-orthogonal) dictio- naries via l1 minimization. In *Proc. Nat. Acad. Sci.*, volume 100, pages 2197–2202, 2003.

[7] J. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with applica- tion to arithmetic complexity and statistics. *Lin. Algebra Appl.*, 18(2):95–138, 1977.

[8] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Infomation Theory*, 47(7):2845–2862, September 2001.

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

[10] Yi Lin Ming Yuan, Ming Yuan and Yi Lin. Model selection and estimation in re- gression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:9–67, 2006.

[11] I.F. Gorodnitsky and B.D. Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *Transaction on Signal Processing*, 45(3):600–616, March 1997.

[12] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[13] D.L. Donoho, E. Elad, and V.Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Infomation Theory*, 52(1):6–18, 2006.

[14] Z. Ben-Haim, Y.C. Eldar, and M. Elad. Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Transactions on Signal Processing*, 58(10):5030 – 5043, October 2010.

[15] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshiran. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[16] Yingying Li and Stanley Osher. Coordinate descent optimization for 1 minimization with application to compressed sensing ; a greedy algorithm. *Inverse Problems and Imaging*, 3:487–503, 2009.

[17] D. D. Lewis A. Genkin and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[18] W. Yin Z. Peng, M. Yan. Parallel and distributed sparse optimization. In *IEEE Asilomar Conference on Signals, Systems, and Computers*, 2013.

[19] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

[20] Donald Goldfarb Wotao Yin, Stanley Osher and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[21] I S Dhillon, P Ravikumar, and A Tewari. Nearest neighbor based greedy coordinate descent. In *The Neural Information Processing Systems (NIPS)*, 2011.

[22] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1 regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 929–936, New York, NY, USA, 2009. ACM.

[23] Léon Bottou and Yann LeCun. On-line learning for very large datasets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

[24] Stephen J. Wright, Robert D. Nowak, and Mário A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, July 2009.

[25] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, November 2004.

[26] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, March 2009.

[27] J. Barzilai and J. M. Borwein. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

[28] Elaine T. Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. Optim*, 19(3):1107–1130, 2008.

[29] R. Nowak M. Figueiredo and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, December 2007.

[30] J. Bobin S. Becker and E. J. Candès. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM J. on Imaging Sciences*, 4(1):1–39, 2009.

[31] Allen Y. Yang, Zihan Zhou, Arvind Ganesh Balasubramanian, Shankar Sastry, and Yi Ma. Fast l1-minimization algorithms for robust face recognition. *IEEE Transactions on Image Processing*, 22(8):3234–46, August 2013.

[32] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, 1989.

[33] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, April 2004.

[34] Rajat Raina Honglak Lee, Alexis Battle and Andrew Y. Ng. NIPS. Efficient sparse coding algorithms. In *NIPS*, pages 801–808, 2006.

[35] R. Glowinski and A. Marrocco. Sur lapproximation par elements finis dordre un, et la resolution par penalisation-dualite dune classe de problemes de dirichlet nonlineaires. *Rev. Francaise dAut. Inf. Rech. Oper.*, R-2:41–76, 1975.

[36] D. Gabay and B. Mercier. A dual algorithm for the solution of non- linear variational problems via finite-element approximations. *Comp. Math. Appl.*, 2:17–40, 1976.

[37] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, January 2011.

[38] John Langford, Alexander J. Smola, and Martin Zinkevich. Slow learners are fast. In *The Neural Information Processing Systems (NIPS)*, 2009.

[39] Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *The Journal of Machine Learning Research*, 11:3183–3234, December 2010.

[40] Danny Bickson Carlos Guestrin Joseph K Bradley, Aapo Kyrola. Parallel coordinate descent for l1-regularized loss minimization. In *International Conference on Machine Learning (ICML)*, June 2011.

[41] Peter Richtárik and Martin Takáč. Efficient serial and parallel coordinate descent methods for huge-scale truss topology design. *Operations Research Proceedings 2011*, pages 27–32, 2011.

[42] Mahantesh Halappanavar David Haglin Chad Scherrer, Ambuj Tewari. Feature clustering for accelerating parallel coordinate descent. In *The Neural Information Processing Systems (NIPS)*, 2012.

[43] P. Richtarik and M. Takac. Distributed coordinate descent method for learning with big data. Technical report, arXiv: 1310.2059, 2013.

[44] Michael Elad, Boaz Matalon, and Michael Zibulevsky. Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. *Applied and Computational Harmonic Analysis*, 23(3):346–367, November.

[45] J.-A. Bazerque G. Mateos and G. B. Giannakis. Distributed sparse linear regression. *EEE Transactions on Signal Processing*, 58(10):5262 – 5276, 2010.

[46] J. H. Husøy K. Engan, S. O. Aase. Method of optimal directions for frame design. In *Proc. ICASSP'99*, pages 2443–244, March 1999.

[47] M. Aharon, M. Elad, and A. Bruckstein. k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[48] Michael Zibulevsky Ron Rubinstein and Michael Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE Transactions on Image Processing*, 58(3):1553 – 1564, 2010.

[49] Jean Ponce Guillermo Sapiro Andrew Zisserman Julien Mairal, Francis Bach. Supervised dictionary learning. In *The Neural Information Processing Systems (NIPS)*, 2008.

[50] D. M. Bradley and J. A. Bagnell. Differential sparse coding. In *The Neural Information Processing Systems (NIPS)*, 2008.

[51] Jean Ponce Guillermo Sapiro Julien Mairal, Francis Bach. Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*, 2009.

[52] Z. Szabo, B. Poczos, and A. Lorincz. Online group-structured dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2865–2872, 2011.

[53] Hao Xu Zhen J. Xiang and Peter J. Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In *The Neural Information Processing Systems (NIPS)*, 2011.

[54] Vikas Sindhwani and Amol Ghoting. Large-scale distributed non-negative sparse coding and sparse dictionary learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 489–497, 2012.

[55] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, December 2004.

[56] R. Peharz, M. Stark, and F. Pernkopf. Sparse nonnegative matrix factorization using l0-constraints. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 83–88, 2010.

[57] Michael Elad Michal Aharon and Alfred M. Bruckstein. K-svd and its non-negative variant for dictionary design. In *Proceedings of the SPIE conference wavelets*, pages 327–339, 2005.

[58] S. Si I. S. Dhillon H.-F. Yu, C.-J. Hsieh. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. i. In *IEEE International Conference on Data Mining(ICDM)*, 2012.

[59] Rainer Gemulla, Erik Nijkamp, Peter J. Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 69–77, New York, NY, USA, 2011. ACM.

[60] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[61] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[62] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

[63] G. Sapiro J. Mairal and M. Elad. Learning multiscale sparse repre- sentations for image and video restoration. *SIAM Interdisciplinary Journal, Multiscale Modeling and Simulation*, 7(1):214–241, April 2008.

[64] Sophia Antipolis France ; Bach F. ; Ponce J. ; Sapiro G. J. Mairal F. Bach J. Ponce G. Sapiro Mairal, J. ; INRIA and A. Zisserman. Non-local sparse models for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[65] P. Chatterjee and P. Milanfar. Clustering-based denoising with locally learned dictionaries. *IEEE Transactions on Image Processing*, 18(7):1438–1451, 2009.

[66] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[67] Jia-Bin Huang Chih-Yuan Yang and Ming-Hsuan Yang. Exploiting self-similarities exploiting self-similarities for single frame super-resolution. In *Proceedings of the 10th Asian conference on Computer vision - Volume Part III*, pages 497–510, 2010.

[68] Michael Elad Roman Zeyde and Matan Protter. On single image scale-up using sparse-representations. In *Proceedings of the 7th international conference on Curves and Surfaces*, pages 711–730, 2012.

[69] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2216 – 2223, 2012.

[70] Lei Shi Guangming Weisheng Dong Dong, Weisheng Zhang, Lei Zhang, Guangming Shi, and XiaolinWu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838 – 1857, 2011.

[71] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[72] Ehsan Elhamifar and Rene Vidal. Robust classification using structured sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011.

[73] Nasser M. Nasrabadi Rama Chellappa Sumit Shekhar, Vishal M. Patel. Back to results joint sparse representation for robust multimodal biometrics recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):113 – 126, 1 2014.

[74] N.H. Nguyen, N.M. Nasrabadi, and T.D. Tran. Robust multi-sensor classification via joint sparse representation. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8, 2011.

[75] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[76] Jian Yang David Zhang Meng Yang, Lei Zhang. Regularized robust coding for face recognition. *IEEE Transactions on Image Processing*, 22(5):1753–1766, 2013.

[77] Meng Yang Lei Zhang and Xiangchu Feng. Sparse representation or collaborative representation : Which helps face recognition ? In *Tenth IEEE International Conference on Computer Vision (ICCV2011)*, 2011.

[78] Yi Ma Allen Y. Yang, Zihan Zhou and S. Shankar Sastry. Towards a robust face recognition system using compressive sensing. In *InterSpeech 2010*, 2010.

[79] E. J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, pages 1433–1452, Madrid, 2006.

[80] Jianchao Yang, Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[81] Liang-Tien Chia Shenghua Gao, Ivor Wai-Hung Tsang and Peilin Zhao. Local features are not lonely – laplacian sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[82] Tong Zhang Kai Yu and Yihong Gong. Nonlinear learning using local coordinate coding. In *The Neural Information Processing Systems (NIPS)*, 2009.

[83] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained liner coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[84] Kai Yu Jianchao Yang and Thomas Huang. Efficient highly over-complete sparse coding using a mixture model. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV 2010, pages 113–126, Berlin, Heidelberg, 2010. Springer-Verlag.

[85] Kai Yu Jianchao Yang and Thomas Huang. Supervised translation-invariant sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[86] John Lafferty Kai Yu, Yuanqing Lin. Learning image representations from the pixel level via hierarchical sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[87] Yun Wang Arthur Szlam Yunlong He, Koray Kavukcuoglu and Yanjun Qi. Unsupervised feature learning by deep sparse coding. In *SIAM 2014 International Conference on Data Mining (SDM2014)*, 2014.

[88] Fengjun Lv Shenghuo Zhu Ming Yang Timothee Cour Kai Yu Yuanqing Lin, Liangliang Cao and Thomas Huang. Large-scale image classification: Fast feature extraction and svm training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011.

[89] L.Daudet R Gribonval M. E. Davies M.D.Plumbley, T. Blumensath. Sparse representations in audio and music: From coding to source separation. In *IEEE*, volume 98, pages 995–1005. IEEE, 2010.

[90] C. Fevotte, B.Torresani, L. Daudet, and S. J. Godsill. Sparse linear regression with structured priors and application to denoising of musical audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):174–185, 2008.

[91] Samer A. Abdallah and Mark D. Plumbley. Sparse coding of music signals. Technical report, Department of Electronic Engineering. King's College London, March 2001.

[92] Samer A. Abdallah and Mark D. Plumbley. Unsupervised analysis of polyphonic unsupervised analysis of polyphonic music by sparse coding. *IEEE Transactions on Neural Networks*, 17(1):179–196, 2006.

[93] Thomas Blumensath Mark D. Plumbley, Samer A. Abdallah and Michael E. Davies. Sparse representations of polyphonic music. *IEEE Transactions on Signal Processing*, 86(3):417–431, 2006.

[94] Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, October 2003.

[95] Michal Genussov and Israel Cohen. Multiple fundamental frequency estimation based on sparse representations in a structured dictionary. *Digital Signal Processing*, 23(1):390–400, 2013.

[96] Yi-Hsuan Yang Cheng-Te Lee and Homer H. Chen. Multipitch estimation of piano music by exemplar-based multipitch estimation of piano music by exemplar-based sparse representation. *IEEE Transactions on Multimedia*, 14(3):608–618, 2012.

[97] Mark D. Plumbley Ken O'Hanlon ? Ken O'Hanlon, Hidehisa Nagano. Structured sparsity for automatic music transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 441 – 444, 2012.

[98] Constantine Kotropoulos Yannis Panagakis and Gonzalo R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In *17th European Signal Processing Conference (EUSIPCO)*, 2009.

[99] T. Virtanen J. F. Gemmeke and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition,. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2067–2080, September 2011.

[100] Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionar. *MIT Press Journals*, 13(4):863–882, 2001.

[101] Tuomas Virtanen. Sound source separation using sparse coding with sound source separation using sparse coding with sound source separation using sparse coding with sound source separation using sparse coding with temporal continuity objective. In *International Computer Music Conference*, 2003.

[102] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3449 – 3456, 2011.

[103] Eric Xing Bin Zhao, Li Fei-Fei. Online detection of unusual events in videos via dynamic sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[104] Yacov Hel-Or Amir Adler, Michael Elad and Ehud Rivlin. Sparse coding with anomaly detection. In *IEEE International Workshop on Machine Learning for Signal Processing*, 2013.

[105] S. P. Kasiviswanathan, G. Cong, P. Melville, and R. D. Lawrence. Novel document detection for massive data streams using distributed dictionary learning. *IBM Journal of Research and Development*, 57(3/4):9:1–9:15, MAY/JULY 2013.