DOI:10.13718/j.cnki.xsxb.2023.04.001

稀疏统计学习及其最新研究进展综述①

张红英, 董珂臻

西安交通大学 数学与统计学院, 西安 710049

摘要: 稀疏性意谓可以仅用少数位于低维子空间的参数(特征变量)近似表示高维空间的复杂物理过程, 是实际应用中普遍存在的性质. 稀疏统计学习旨在探索高维数据的稀疏性, 并进行统计建模和推断. 文章综述了基于回归分析的稀疏统计学习模型及其最新研究进展. 主要介绍了各类带有凸或非凸正则项的稀疏回归模型, 特别是 $L_{\frac{1}{2}}$-正则化框架的算法和应用. 近 10 年来, 深度学习取得革命性进展, 结合传统稀疏统计学习模型与深度神经网络的研究逐渐受到了广泛的关注. 文章主要介绍了基于稀疏建模的深度学习方法和数据驱动的稀疏统计分析方法, 前者包括深度网络展开等, 后者则包括深度哈希学习及深度典型相关分析. 最后, 文章进行了总结, 并展望了未来可能的研究方向.

关 键 词: 稀疏性; 正则化框架; 正则项; $L_{\frac{1}{2}}$-正则化框架; 深度学习; 深度网络展开

中图分类号: TP181 **文献标志码**: A **文章编号**: 1000-5471(2023)04-0001-12

A Review of Sparse Statistical Learning and Its Recent Research Progress

ZHANG Hongying, DONG Kezhen

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China

Abstract: Sparsity means that complex physical processes in high-dimensional spaces can be approximated by only a few parameters (characteristic variables) located in low-dimensional subspaces, and is a prevalent property in practical applications. Sparse statistical learning aims to explore the sparsity of high-dimensional data and to perform statistical modeling and inference. The article reviews the sparse statistical learning models with a focus on regression analysis and its recent research progress. It mainly introduces various types of sparse regression models with convex or non-convex regularization terms, especially the algorithms and applications of $L_{\frac{1}{2}}$-regularization framework. In the last decade, deep learning has made revolutionary progress, and the research combining traditional sparse statistical learning models with deep neural networks has gradually received widespread attention. The article mainly introduces the deep learning methods based on sparse modeling and data-driven sparse statistical analysis methods, the former including deep unfolding networks and so on, and the latter including deep hash learning and deep canoni-

① 收稿日期: 2022-06-17
基金项目: 国家自然科学基金面上项目(12171386, 11671007).
作者简介: 张红英, 教授, 博士研究生导师, 主要从事人工智能的数学基础、认知不确定大数据分析、基于信息理论的机器学习和统计学习等研究.

cal correlation analysis. Finally, the article concludes with a summary and looks at possible future research directions.

Key words: sparsity; regularization framework; regularization terms; $L_{\frac{1}{2}}$-regularization framework; deep learning; deep unfolding networks

稀疏性是高维空间中信号或者数据的普遍内蕴属性，意谓可以仅用少数位于低维子空间的参数(特征变量)近似表示高维空间的复杂物理过程. 例如，在压缩感知领域，图像可用小波基线性表示并得到表示系数，保留较大系数，对较小系数赋值为0，可得到图像的近似表示. 通过此近似表示对图像进行复原，基本可以恢复原图[1]. 因此，图像可通过少数系数(特征变量)近似线性表示，称为图像的稀疏表示. 稀疏性广泛存在于高维数据特征选择、稀疏信号恢复以及众多其他问题[2-4]之中. 这些问题具有共同特点：在数据生成过程中，特征变量的数量大于采样数量. 具有这样特点的问题被称为高维统计分析问题. 稀疏统计学习正是处理这类问题的有效方法.

变量选择是稀疏统计学习的核心问题之一. 直觉的处理方法是在所有变量子集中，选出使模型拟合性能最好的变量子集，即最优子集选择方法. 然而，若特征变量的数量稍大，最优子集选择方法的计算消耗便十分巨大，因此，基于统计学习正则化框架的变量选择方法逐渐进入研究者视野. Lasso模型即是其中的代表性方法之一. Lasso模型可看作最优子集选择方法的最紧凸近似，在计算效率和变量选择能力上优势显著，因此迅速受到广泛关注. 大量基于Lasso惩罚的模型被提出并得到应用，同时涌现出众多基于改进Lasso惩罚得到的广义Lasso模型，例如自适应Lasso模型、弹性网模型、组Lasso模型、稀疏组Lasso模型、融合Lasso模型以及非参模型中的稀疏加法模型等. 然而，Lasso模型有变量选择不一致及不具有Oracle性质等缺陷. 为了克服Lasso模型的缺陷，诸如$L_{\frac{1}{2}}$-惩罚模型、SCAD模型及MCP模型等基于非凸惩罚的模型被提出. 此类模型具有Oracle性质.

近10年来，深度学习在人脸识别、语音处理和文本分析等领域获得空前成功，在众多任务上均达到接近甚至超越人类的性能表现. 然而，大量的网络参数会带来训练代价巨大、容易过拟合等问题，网络的黑箱特性也导致模型可解释性较弱. 稀疏统计学习固有的约简特性和强可解释性，提供了解决上述问题的一种可能. 将稀疏统计学习与深度学习结合，成为受到广泛关注的热点领域. 相关研究大致可以分为基于稀疏建模的深度学习方法和数据驱动的稀疏统计分析方法. 前者通过稀疏建模的思想与方法，进行深度神经网络架构和算法的设计. 后者则利用深度神经网络的强大表示能力，通过深度神经网络学习特征表示，并应用于稀疏统计学习模型. 越来越多的研究集中于该领域，并获得了令人欣喜的成果.

本文从经典的稀疏统计学习方法出发，回顾经典的稀疏回归分析方法，并对深度神经网络与稀疏统计学习相结合的研究进行简要综述.

1 稀疏回归分析

回归分析方法作为统计数据分析强有力的工具之一，一直是统计学的研究热点，同时也被广泛应用于自然科学及社会科学的各个领域. 回归分析旨在描述输出变量与特征变量的关系，并进行统计建模和推断. 近年来，随着计算与数据采集能力的持续提高，高维数据逐渐成为回归分析的重要对象. 为得到正定解，在高维数据回归分析中，一般使用基于稀疏性假设的正则化框架. 本节将在正则化框架下，综述稀疏回归分析的各类模型.

1.1 正则化框架

假设存在度量空间$Z=X\times Y$，其中$X\subseteq\mathbb{R}^p$是特征空间，Y是输出空间，并在度量空间Z中独立同分布地抽取随机样本$\{Z_i=(X_i,y_i)\}_{i=1}^n$. 假设输入与输出之间存在函数关系$y=f(X)$，其中$f\in H$，$H$为假设空间. 回归分析的重要目标之一是通过有限的随机样本估计特征空间与输出空间的函数关系f. 为此，一类被广泛应用的重要方法将问题建模为如下模型：

$$\min_{f\in H}\left\{\frac{1}{n}\sum_{i=1}^{n}L(f(X_i),y_i)+\lambda P(f)\right\} \tag{1}$$

模型(1) 称为统计学习的正则化框架，其中 $L(\cdot)$ 是损失函数，$P(\cdot)$ 是正则项，λ 是正则化参数. 特定的正则化框架会使模型的解具有正定性、光滑性和稀疏性等性质，从而提高了模型的精度和可解释性. 不同的方法常根据不同的先验信息选择不同的损失函数和正则项.

1.2　稀疏回归分析中的变量选择

稀疏性假设是针对高维数据进行统计建模和分析的重要手段. 稀疏性假设意指在统计模型中，仅少数特征变量对输出产生重要影响. 基于稀疏性假设，稀疏回归分析旨在基于回归分析框架对高维数据进行统计建模、分析和推断.

因仅少数特征变量产生重要影响，变量选择便成为稀疏回归分析中的关键问题. 最优子集选择[5] 则是最直接的变量选择方法.

对于多元线性回归模型

$$y=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon \tag{2}$$

其中 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)$，$\boldsymbol{\beta}=(\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$，$\varepsilon$ 是随机噪声. 最优子集选择在所有包含 $m(0\leqslant m\leqslant p)$ 个变量的模型中，选择拟合效果最好(残差最小) 的一个.

最优子集的选择可以被纳入统计学习的正则化框架，可以被看作是基于 L_0-正则项的最小二乘模型. 假设 $\|\boldsymbol{\beta}\|_0$ 是向量 $\boldsymbol{\beta}$ 的 L_0-范数，其中 L_0-范数表示向量 $\boldsymbol{\beta}$ 中的非零元素个数，则基于 L_0-惩罚的最小二乘估计准则为

$$\min_{\boldsymbol{\beta}}\{\|y-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_0\} \tag{3}$$

当 $\|\boldsymbol{\beta}\|_0=m$ 时，模型(3) 的解等价于最优子集选择的结果. 求解模型(3) 是 NP-难问题[6]，已知算法均无法在多项式时间内有效解决. 因此，求解过程通常需要进行近似. 值得指出的是，一些准则例如 AIC 准则[7]、BIC 准则[8]、HQIC 准则[9] 等也是 L_0-正则化模型.

1.3　基于凸正则项的稀疏回归分析

求解模型(3) 的 NP-难问题，一个重要方法就是利用正则化框架进行松弛处理.

Tikhonov 正则化方法通过利用控制函数光滑性的惩罚项解决积分方程不可解或者有无穷多解的问题，是一种标准的求解非适定积分方程的方法. 以求解逆问题的观点来看，岭回归模型可以认为是 Tikhonov 正则化方法的特例[10]，其形式为

$$\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\|y-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2\right\} \tag{4}$$

模型(4) 用 L_2-正则项代替模型(3) 中的 L_0-正则项，具有解析解，并具有收缩性质，即迫使参数 $\boldsymbol{\beta}$ 的所有元素趋向于 0，却不恰等于 0. 因此，模型(4) 不具有变量选择性质，且在参数较大时会带来偏差.

文献[11] 提出了基于线性回归模型的非负绞刑模型. 该模型可看作 3 个步骤：

(i) 设定初始估计量 $\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p$；

(ii) 求解模型

$$\min_{\boldsymbol{c}\in\mathbb{R}^p}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}c_i x_{ij}\tilde{\beta}_j\right)^2+\lambda\|\boldsymbol{c}\|_1\right\} \tag{5}$$

使得 $\boldsymbol{c}\geqslant 0$，得到最优解 $\hat{\boldsymbol{c}}$；

(iii) 得到估计 $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{c}}\odot\tilde{\boldsymbol{\beta}}$，其中 $\odot$ 表示 Hadamard 积.

非负绞刑模型具有变量选择性质，并且可以得到比最优子集选择和逐步选择方法更加稳定的解.

受到非负绞刑模型的启发，文献[12] 提出了 Lasso 模型. 该模型使用 L_1-正则项替代最优子集选择中的 L_0-正则项. 几乎在同一时期，L_1-正则项同样被用于信号处理领域的基追踪方法[13]. 基于 L_1-正则项的 Lasso 模型具有形式

$$\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\|y-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\right\} \tag{6}$$

Lasso 惩罚具有变量选择性质，且是 L_0-范数的最紧凸松弛，在特定条件下，两者的解完全等价. 大量的研究显示，基于 L_1-正则项的 Lasso 模型具有强可解释性、统计有效性[14] 和计算高效性等良好性质，因

此得到了广泛关注和应用.

Lasso模型在变量选择过程中,通过对全部特征变量的系数施加相同程度的惩罚进行系数收缩,以达到将与响应变量无关的冗余变量压缩为0的目的.然而,这会使得与响应变量相关的目标变量的系数也受到相同程度的压缩,导致回归系数的估计是有偏的.

为得到无偏或者近似无偏的估计,文献[15]提出了自适应Lasso模型.自适应Lasso模型具有形式

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \| y - \boldsymbol{X\beta} \|_2^2 + \lambda \| \boldsymbol{w\beta} \|_1 \right\} \tag{7}$$

其中,$\boldsymbol{w} \in \mathbb{R}^p$ 是已知的权重向量.自适应Lasso模型采用重新加权的 L_1-范数,能够修正Lasso模型的过度估计,并且具有Oracle性质.

Lasso模型的另一个缺陷是无法妥善处理特征变量间具有高相关性的数据.当一组特征变量两两之间相关性很高时,Lasso模型倾向于只选择其中任意一个;当 $n < p$ 时,Lasso模型至多只能选择 n 个变量;当 $n > p$ 且特征变量间有强相关性时,Lasso模型的性能逊于岭回归模型.为了克服此缺陷,文献[16]提出了弹性网模型.该模型的惩罚项是Lasso惩罚与岭回归惩罚的凸组合,具体形式为

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \| y - \boldsymbol{X\beta} \|_2^2 + \lambda_1 \| \boldsymbol{\beta} \|_1 + \lambda_2 \| \boldsymbol{\beta} \|_2^2 \right\} \tag{8}$$

当特征变量高相关时,弹性网模型会使这些变量的系数趋向于相同,因此,弹性网模型可以选到全部相关的特征变量.

另一种研究特征变量间相关性的方法是利用其组结构,将特征变量分组,研究不同组的特征变量与输出之间的关系.组Lasso模型[17]就用来解决此类问题,其形式为

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \| y - \sum_{j=1}^{J} x_j \beta_j \|_2^2 + \lambda \sum_{j=1}^{J} \| \beta_j \|_{K_j} \right\} \tag{9}$$

其中,p 个特征变量被分为 J 组,$\| \boldsymbol{\beta} \|_{K_j} = (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{K}_j \boldsymbol{\beta})^{\frac{1}{2}}$.当 β_j 为单个元素且 $\boldsymbol{K}_j$ 为单位矩阵时,模型(9)退化为Lasso模型.

当某个组在组Lasso模型拟合中被选中时,该组中的所有系数均不为0,因此无法处理组内个别目标变量的回归系数不为0的情况.为了实现组内稀疏性,稀疏组Lasso模型[18-19]对组Lasso模型进行了改进,其形式为

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \| y - \sum_{j=1}^{J} x_j \beta_j \|_2^2 + \lambda_1 \| \boldsymbol{\beta} \|_1 + \lambda_2 \sum_{j=1}^{J} \| \beta_j \|_{K_j} \right\} \tag{10}$$

其中 $\lambda_1, \lambda_2 \geqslant 0$.模型(10)可以同时实现组间与组内稀疏性.当 $\lambda_1 = 0$ 时,模型(10)退化为组Lasso模型;当 $\lambda_2 = 0$ 时,模型(10)退化为Lasso模型.

Lasso模型无法处理连续变量数据.为克服这个缺陷,融合Lasso模型[20]对Lasso模型进行了扩展,其具体形式为

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \| y - \boldsymbol{X\beta} \|_2^2 + \lambda_1 \| \boldsymbol{\beta} \|_1 + \lambda_2 \| \boldsymbol{B\beta} \|_1 \right\} \tag{11}$$

其中 $\lambda_1, \lambda_2 \geqslant 0$, $\boldsymbol{B\beta} = [\beta_1 - \beta_2, \beta_2 - \beta_3, \cdots, \beta_{p-1} - \beta_p]^{\mathrm{T}}$.模型(11)通过促使相邻系数趋于相同以保证获得稀疏解以及数据的局部连续性.

除了上述参数回归模型之外,一些基于稀疏正则化框架的非参数模型也得到了广泛研究.非参数回归往往受到"维数灾难"的困扰,因此近似方法对于该类模型至关重要.加法模型正是此类近似方法,该模型为

$$y = \sum_{j=1}^{p} f_j(x_j) + \varepsilon \tag{12}$$

通过在加法模型上应用Lasso惩罚,文献[21]提出了稀疏加法模型

$$\min_{\boldsymbol{\beta}} \left\{ \| y - \sum_{j=1}^{p} \beta_j f_j(x_j) \|_2^2 + \lambda \sum_{j=1}^{p} \| \beta_j \|_1 \right\} \tag{13}$$

任何非参数方法均可以用于拟合模型(13).Backfitting算法[22]同样适用于计算模型(13).

1.4　基于非凸正则项的稀疏回归分析

L_0 -正则化模型中 L_0 -范数是向量中非零元素的个数，为非凸非连续函数. 因此，L_0 -正则化模型是基于非凸正则项的稀疏回归模型.

文献[23-25]发现 Lasso 估计需要在特定条件下才具有较好的变量估计和选择特性，且即使在这些条件下，Lasso 估计仍存在偏差. 为克服上述缺陷，可采用非凸的 $0<q<1$ 时的 L_q -正则化框架，其模型为

$$\min_{\boldsymbol{\beta}} \{ \| y-\boldsymbol{X\beta} \|_2^2+\lambda \| \boldsymbol{\beta} \|_q^q \} \tag{14}$$

其中，$\| \boldsymbol{\beta} \|_q=\left(\sum_{i=1}^{N} | \beta_i |^q\right)^{\frac{1}{q}}$.

文献[26]提出了 bridge 回归模型

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \| y-\boldsymbol{X\beta} \|_2^2+\sum_{i=1}^{p}\lambda | \beta_i |^{\gamma} \right\} \tag{15}$$

其中 $0<\gamma$. 模型(15)搭起了最优子集选择与岭回归之间的桥梁. 当 $0<\gamma<1$ 时，正则项是非凸的，其等价于 $0<q<1$ 时的 L_q -正则化框架.

大量研究[27-29]发现，相较于 L_1 -正则化框架，$0<q<1$ 时的 L_q -正则化框架有着需要更少的采样数以及变量选择能力更强等显著优势. 然而，当 $0<q<1$ 时，L_q -正则化框架是非凸非光滑非李普希兹的优化问题，理论上难以直接求解. 同时，如何选取 q 也是一个重要问题. 文献[28,30-31]揭示了 $q=\frac{1}{2}$ 时的 $L_{\frac{1}{2}}$ -正则化框架在 $0<q<1$ 时的 L_q -正则化框架中具有代表性地位，并在实际应用中表现突出. $L_{\frac{1}{2}}$ -正则化框架具有形式

$$\min_{\boldsymbol{\beta}} \{ \| y-\boldsymbol{X\beta} \|_2^2+\lambda \| \boldsymbol{\beta} \|_{\frac{1}{2}}^{\frac{1}{2}} \} \tag{16}$$

$L_{\frac{1}{2}}$ -正则化框架的求解是困难的问题. 文献[32]证明了 $L_{\frac{1}{2}}$ -正则化框架的阈值表示定理，并据此提出了求解 $L_{\frac{1}{2}}$ -正则化问题的 Half 阈值迭代算法. 该定理证明了 $L_{\frac{1}{2}}$ 正则化问题的解满足不动点阈值表示

$$\boldsymbol{\beta}=T_{\lambda\mu,\frac{1}{2}}(\boldsymbol{\beta}+\mu\boldsymbol{X}^{\mathrm{T}}(y-\boldsymbol{X\beta})) \tag{17}$$

其中，$\mu \in \left(0, \frac{1}{\| \boldsymbol{X} \|^2}\right]$ 是任意正实数. $T_{\lambda\mu,\frac{1}{2}}$ 是 Half 阈值函数，其形式为

$$[T_{\lambda\mu,\frac{1}{2}}(y)]_j=\begin{cases} f_{\lambda\mu,\frac{1}{2}}(y_j) & | y_j |>y^* \\ 0 & \text{否则} \end{cases} \tag{18}$$

其中

$$f_{\lambda,\frac{1}{2}}(y_j)=\frac{2}{3}y_j\left(1+\cos\left(\frac{2\pi}{3}-\frac{2\varphi_{\lambda}(y_j)}{3}\right)\right)$$

且

$$\varphi_{\lambda}(y_j)=\arccos\left(\frac{\lambda}{8}\left(\frac{| y_j |}{3}\right)^{-\frac{3}{2}}\right)$$

$y^*=\frac{\sqrt[3]{54}}{4}(\lambda\mu)^{\frac{2}{3}}$ 是阈值. 文献[32]同样给出了 Half 阈值迭代算法的正则化参数选择策略以及收敛性分析.

$L_{\frac{1}{2}}$ -正则化框架以其良好的性质被广泛应用于压缩感知[33-34]、矩阵分解、图像复原[35]及高光谱图像等领域.

例如在矩阵的稀疏-低秩分解领域，文献[36]将 $L_{\frac{1}{2}}$ -惩罚引入稀疏-低秩矩阵分解问题，提出了模型

$$\min_{\boldsymbol{A},\boldsymbol{E}\in\mathbb{R}^{m\times n}} \{ \| \boldsymbol{A} \|_{S_{\frac{1}{2}}}^{\frac{1}{2}}+\lambda \| \boldsymbol{E} \|_{1_a}^{a} \} \tag{19}$$

使得 $\| \boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E} \|_F \leqslant \delta$. 其中，$\| \boldsymbol{A} \|_{S_{\frac{1}{2}}}=\left(\sum_{i=1}^{r}\sigma_i^{\frac{1}{2}}\right)^2$ 表示矩阵 $\boldsymbol{A}$ 的所有奇异值构成向量的 $L_{\frac{1}{2}}$ -范数，$\| \boldsymbol{E} \|_{1_a}^{a}=\left(\sum_{i=1}^{m}\sum_{j=1}^{n} | E_{ij} |^a\right)^{\frac{1}{a}}$ 表示矩阵 $\boldsymbol{E}$ 拉直向量的 l_a -范数，参数 a 可根据不同的噪声水平 δ 选取不同

的值(1或者$\frac{1}{2}$). 模型优化过程基于ADMM算法[37]思想,将作用于向量的Soft[38]和Half阈值算子推广到了矩阵情形,设计了稳健且高效的算法.

类似地,文献[39]尝试利用基于$L_{\frac{1}{2}}$-惩罚的矩阵低秩表示模型解决高光谱图像分类问题,提出了模型

$$\min_{\boldsymbol{Z},\boldsymbol{E}}\{\|\boldsymbol{Z}\|_{\frac{1}{2}}^{\frac{1}{2}}+\lambda\|\boldsymbol{E}\|_{2,1}\} \tag{20}$$

使得$\boldsymbol{D}=\boldsymbol{D}\boldsymbol{Z}+\boldsymbol{E}$. 模型(20)利用增广拉格朗日乘子方法(ALM)[40]和Half阈值算子进行求解.

受到文献[32]的启发,文献[41]仔细考察了基于矩阵的$L_{\frac{1}{2}}$-正则化框架模型

$$\min_{\boldsymbol{D}\in\mathbb{R}^{m\times n}}\{\|\boldsymbol{X}\boldsymbol{D}-\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{D}\|_{\frac{1}{2}}^{\frac{1}{2}}\} \tag{21}$$

并提出了求解的迭代算法,同时证明了其收敛性.

基于文献[41]的结果,文献[42]通过结合TV正则项[43]和$L_{\frac{1}{2}}$-正则项对矩阵稀疏-低秩分解问题进行了建模,并应用于运动目标检测问题,提出了模型

$$\min_{\boldsymbol{A},\boldsymbol{E}\in\mathbb{R}^{m\times n}}\{\|\boldsymbol{A}\|_*+\lambda\Omega(\boldsymbol{E})+\mu\|\boldsymbol{E}\|_{l_{\frac{1}{2}}}^{\frac{1}{2}}\} \tag{22}$$

使得$\|\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}\|_F^2\leqslant\varepsilon$, $\operatorname{rank}(\boldsymbol{A})\leqslant r$. 其中

$$\Omega(\boldsymbol{E})=\sum_{k=1}^{p}\|\boldsymbol{E}_k\|_{\mathrm{TV}}$$

$$\|\boldsymbol{E}_k\|_{\mathrm{TV}}=\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{(\boldsymbol{E}_{k_h}(i,j))^2+(\boldsymbol{E}_{K_v}(i,j))^2}+\sum_{i=1}^{m-1}|\boldsymbol{E}_{K_v}(i,n)|+\sum_{j=1}^{n-1}|\boldsymbol{E}_{K_h}(m,j)|$$

$\boldsymbol{E}_{K_h}$和$\boldsymbol{E}_{k_v}$分别表示水平方向和竖直方向的运算.

模型(22)针对2维矩阵的模型,文献[44]将模型(22)推广至3维张量的情形,结合TTV正则项[45]和$L_{\frac{1}{2}}$-正则项提出了基于张量框架的模型

$$\min_{\boldsymbol{S},\boldsymbol{T},\boldsymbol{Z},\boldsymbol{C}}\{\|\boldsymbol{S}\|_{\odot}+\lambda_1\|\boldsymbol{T}\|_{l_{\frac{1}{2}}}^{\frac{1}{2}}+\lambda_2\|\boldsymbol{Z}\|_{l_{\frac{1}{2}}}^{\frac{1}{2}}+\lambda_3\|\boldsymbol{C}\|_{\mathrm{TTV}}\} \tag{23}$$

使得$\boldsymbol{R}=\boldsymbol{S}+\boldsymbol{T}$, $\boldsymbol{T}=\boldsymbol{Z}+\boldsymbol{C}$. 其中

$$\|\boldsymbol{C}\|_{\mathrm{TTV}}=\sum_{x=1}^{p-1}|\boldsymbol{C}(x,y,z)-\boldsymbol{C}(x+1,y,z)|+\sum_{y=1}^{q-1}|\boldsymbol{C}(x,y,z)-\boldsymbol{C}(x,y+1,z)|+$$

$$\sum_{z=1}^{n-1}|\boldsymbol{C}(x,y,z)-\boldsymbol{C}(x,y,z+1)|$$

$\odot$是张量核范数[46],表示张量秩的最紧凸松弛. 模型(23)通过基于交替方向最小化(ADM)[47]的增广拉格朗日乘子法(ALM)进行优化求解[40].

鉴于非凸正则项的显著优势,除$0<q<1$时的L_q-正则化框架以外的大量非凸正则项同样得到了广泛研究与应用. 文献[48]提出了一个好的正则项产生的估计需要同时具备稀疏性、无偏性和连续性. 稀疏性保证所产生的估计具有变量选择性质,且可以降低模型的复杂度;无偏性保证估计是近似无偏的,以降低模型的偏差;连续性保证模型性能关于数据的稳定性. L_q-正则化框架所产生的估计均无法同时满足以下3个性质:当$q>1$时,L_q-正则项不满足稀疏性;L_1-正则项不满足无偏性;当$0\leqslant q<1$时,L_q-正则项不满足连续性. 基于此,文献[48]提出了SCAD模型. SCAD模型的形式为

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{2}\|y-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{i=1}^{p}p_\lambda(\beta_j)\right\} \tag{24}$$

其中,$p_\lambda(\beta_i)$的形式为

$$p_\lambda(\beta_i)=\begin{cases}\lambda|\beta_i| & |\beta_i|<\lambda\\ \dfrac{-\lambda^2+2a\lambda|\beta_i|-\beta_i^2}{2(a-1)} & \lambda\leqslant|\beta_i|\leqslant a\lambda\\ \dfrac{(a+1)\lambda^2}{2} & |\beta_i|>a\lambda\end{cases} \tag{25}$$

其中 $\lambda \geqslant 0$，$a > 2$. SCAD 正则项可以产生具有 Oracle 性质的估计.

文献[49]提出了极小极大凹正则项(MCP)，其形式为

$$p_\lambda(\beta_i)=\begin{cases}\lambda \mid \beta_i \mid -\dfrac{\beta_i^2}{2a} & \mid \beta_i \mid \leqslant a\lambda \\ \dfrac{a\lambda^2}{2} & \mid \beta_i \mid > a\lambda\end{cases} \tag{26}$$

其中 $a > 1$. 模型(26)理论上近似无偏，且具有 Oracle 性质.

SCAD 和 MCP 均为 Folded Concave 惩罚函数，分别是软阈值方法和硬阈值方法的拓展. SCAD 是连续的，但 MCP 不连续.

2　深度稀疏统计分析

过去10年，深度神经网络的研究取得了空前的成功，尤其在图像、语音、文本等任务上表现出色. 深度神经网络研究的成功极大地拓宽了处理高维数据方法的边界，提高了处理能力. 然而，现代深度神经网络在训练和应用中通常被当作“黑箱”，其内部原理依然不清晰，可解释性较差，因而无法严格保证模型性能. 另外，现代深度神经网络参数量巨大，训练过程需要大量训练数据. 这使得深度神经网络训练过程中的计算消耗巨大，常需要庞大的计算资源支持，效率较低.

作为处理高维数据的经典方法，稀疏统计分析方法依据统计理论和不同先验信息建模，通常有较强的可解释性. 同时，稀疏统计分析方法并不依靠大量训练数据，求解过程也仅需要少量迭代便能达到较好性能，因而计算消耗较小，效率较高. 如何将稀疏统计分析方法与深度神经网络结合起来，使模型兼具两种方法优点，逐渐成为广受关注的热点问题.

已有研究大致可以分为两个方向：基于稀疏建模的深度学习方法和基于数据驱动的稀疏统计分析方法. 本节将针对这两方面的相关方法进行综述.

2.1　基于稀疏建模的深度学习方法

基于稀疏建模的深度学习方法通过稀疏统计分析方法进行数据建模，并据此进行深度神经网络架构和算法的设计. 此类方法通常包含深度神经网络展开、神经网络剪枝、神经网络架构搜索等主题. 本文以深度神经网络展开为例进行介绍.

深度神经网络展开是基于稀疏建模的深度学习方法的典型代表. 深度神经网络的架构通常需要交替地进行线性和非线性的变换，其中非线性变换借由 ReLU 等激活函数完成. 此类结构与稀疏统计学习模型中的阈值迭代算法十分类似，其中阈值算子可以看作激活函数. 鉴于这种联系，将稀疏编码算法展开成为神经网络的深度神经网络展开方法逐渐受到关注.

早期的深度网络展开方法可以追溯到文献[50]的工作，为了提高稀疏编码算法的计算效率，提出了一种端到端的学习方法(LISTA). 该算法将求解稀疏编码问题的 ISTA 迭代算法[51]的每一步迭代看作循环神经网络的一层，由此得到一个多层循环神经网络，通过学习参数，自动地学得字典和稀疏编码.

稀疏编码问题的目标是求解模型(27)的稀疏编码：

$$\min_{\boldsymbol{x}\in\mathbb{R}^m}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{W}\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{x}\|_1 \tag{27}$$

其中 $\lambda > 0$，$\boldsymbol{W}\in\mathbb{R}^{n\times m}$ 是过完备字典. ISTA 迭代算法是求解稀疏编码模型(27)的常用方法之一. 其迭代过程为

$$\boldsymbol{x}^{l+1}=S_\lambda\left\{\left(\boldsymbol{I}-\frac{1}{\mu}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\right)\boldsymbol{x}^l+\frac{1}{\mu}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{y}\right\}\qquad l=0,1,\cdots \tag{28}$$

其中 S_λ 是逐元素的软阈值算子，其在每个元素上定义为

$$S_\lambda(\boldsymbol{x})=\mathrm{sign}(\boldsymbol{x})\cdot\max\{\|\boldsymbol{x}\|-\lambda,0\} \tag{29}$$

若令 $\boldsymbol{W}_t=\boldsymbol{I}-\left(\frac{1}{\mu}\right)\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}$，$\boldsymbol{W}_e=\left(\frac{1}{\mu}\right)\boldsymbol{W}^{\mathrm{T}}$，则此时公式(28)可改写为

$$\boldsymbol{x}^{l+1}=S_\lambda\{\boldsymbol{W}_t\boldsymbol{x}^l+\boldsymbol{W}_e\boldsymbol{y}\}\qquad l=0,1,\cdots \tag{30}$$

仔细观察公式(30)可以发现，ISTA迭代算法的每一步迭代中，输入 $\boldsymbol{x}^l$ 经过线性变换和软阈值算子，得到新的 $\boldsymbol{x}^{l+1}$. 这可以看作深度神经网络中的一层，其中软阈值算子对应于神经网络中的激活函数. 执行 L 步迭代相当于连接了 L 层的深度神经网络. 基于此，LISTA将ISTA展开为深度神经网络，其训练损失函数为

$$L(\boldsymbol{W}_t, \boldsymbol{W}_e, \lambda) = \frac{1}{N}\sum_{n=1}^{N} \| \hat{\boldsymbol{x}}^n(\boldsymbol{y}^n; \boldsymbol{W}_t, \boldsymbol{W}_e, \lambda) - \boldsymbol{x}^{*n} \|_2^2 \tag{31}$$

其中，$\hat{\boldsymbol{x}}^n(\boldsymbol{y}^n; \boldsymbol{W}_t, \boldsymbol{W}_e, \lambda)$ 是网络输出的对 $\boldsymbol{y}^n$ 稀疏编码的预测值，$\boldsymbol{x}^{*n}$ 是稀疏编码的真实值.

LISTA通过将稀疏编码算法展开为深度神经网络，以一种可学习的方式获得了稀疏编码问题的解. 该模型基于稀疏编码算法，有着强可解释性. 同时，该模型在计算效率上有着显著的优势. 实验表明，在达到同一精度的条件下，该模型比某些经典ISTA方法快将近20倍.

此外，其他针对稀疏编码问题的算法同样可以被展开为深度神经网络. 例如，文献[52-54]将ADMM算法展开为神经网络，文献[55-57]将近端梯度下降算法展开为深度神经网络，均获得了不错的表现.

除了深度神经网络展开外，稀疏统计学习方法在深度学习的其他方面也有着广泛的应用. 例如，神经网络正则化方法Dropout[58]因其可诱导核范数[59-61]而可被看作探索网络稀疏结构的方式之一；深度神经网络的初始化权重方法[62]、特征标准化方法[63]等训练方法看作稀疏信号恢复或低秩矩阵恢复算法中的等距约束性质[64]，从而保证模型的性能表现.

2.2 基于数据驱动的稀疏统计分析方法

实际应用中，稀疏统计分析方法常依赖于手工得到的低阶特征，表示能力有限. 特征表示的好坏往往对模型性能有着重要影响. 基于数据驱动的稀疏统计分析方法通常建立在传统稀疏统计分析模型基础之上，利用深度神经网络强大的特征表示能力，学习数据的高阶特征，并应用于稀疏统计分析方法之中，以提升模型性能.

深度哈希学习便是数据驱动的稀疏回归分析模型的应用之一. 文献[65]提出了深度语义排序模型(DSRH)，将深度卷积神经网络整合到哈希函数中，共同学习特征表示及哈希函数，并保持特征表示与哈希编码之间的相似性，摆脱了手工特征语义表示能力的限制. 同时，该方法利用编码多层次相似度信息的排序表来指导深度哈希函数的学习. 文献[66]提出了深度监督哈希模型(DSH)，该模型基于卷积神经网络框架设计，将成对的图像(相似或者不相似)作为训练输入以学习近似离散的二元哈希编码表示.

相比于含有两阶段过程的深度哈希学习方法，端到端的深度哈希学习方法以其能大幅提高所学哈希编码的表示能力而受到广泛关注. 文献[67]将卷积神经网络引入哈希学习方法，提出了一种深度监督哈希学习方法(DPSH). 模型首先通过网络学习图像的特征表示，然后将此特征表示通过哈希函数映射为哈希编码. 模型以端到端的方式，通过衡量成对标签相似性的损失函数同时学习特征表示和哈希编码. 为了进一步探索标签信息，文献[68]在DPSH基础上增加了一个判别项用以更新二值编码. 文献[69]利用锚点图设计出深度监督哈希学习方法(DAGH)，可以更加高效地获得哈希编码. 模型通过构建样本锚点子集，并建立锚点与哈希编码之间联系的方式达到提高计算效率的目的.

深度典型相关分析也是数据驱动的稀疏多元分析方法的代表之一. 文献[70]提出了早期的深度典型相关分析方法(Deep CCA). 该方法先用深度神经网络分别求出两个视图的投影向量，然后通过最大化两个投影向量的相关性进行求解. Deep CCA在训练过程中需要将全部训练数据作为一个批次，因此不能应对大规模的数据. 为解决该问题，文献[71]提出了随机Deep CCA(SDCCA). 该模型将神经网络参数训练嵌入交替最小二乘方法，以适应小批次随机优化. 文献[72]针对多模态数据，提出了基于深度典型相关分析的处理方法DCCA. 该模型利用深度全连接网络学习文本数据的特征，并利用卷积神经网络(CNN)学习图像数据的特征. 随后两个模态的数据被当作两个视图的数据矩阵输入典型相关分析框架. 深度自编码器同样被用于典型相关分析. 文献[73]基于CCA框架提出了深度典型相关自编码器(DCCAE)，将典型相关分析与深度自编码器进行结合，达到了更好的性能. 文献[74]提出了相关神经网络(CorrNet)以进一步描述重构误差. 该模型可以利用已存在的一个视图准确恢复另一个视图.

3 总结

大数据时代，作为传统统计学习经典方法的稀疏统计学习，在高维数据处理领域发挥着举足轻重的作

用，基于稀疏假设的正则化框架带来了大量高效的高维数据处理方法. 同时，随着深度学习的革命性进展，结合稀疏统计学习与深度神经网络以兼取两种方法优点的研究也日趋受到重视. 本文综述了稀疏统计学习中的经典模型，简要介绍了传统稀疏统计学习与现代深度学习相结合的研究进展. 然而，目前针对此类结合的研究还有巨大的探索空间，接下来，对未来研究方向提出一些展望：

1) 由于传统优化理论与算法的局限性，目前稀疏统计学习的研究多集中于求解凸目标函数. 但实际应用中频繁遇到损失函数和正则项非凸的情形. 同时，非凸正则项通常具有更好的统计性质. 因此，对于含有非凸损失和非凸正则项的模型，包括算法的设计与收敛性的证明，都是值得进一步研究的方向.

2) 由于传统的稀疏统计学习方法通常基于最小二乘损失，其数据服从高斯分布. 然而，现实应用中数据常常并不服从高斯分布. 同时，高斯分布对异常点敏感的特性也限制了它的应用. 因此，探索基于更加鲁棒的损失函数的稀疏统计学习方法，例如基于分位数回归的稀疏统计学习方法等也值得进一步的研究.

3) 深度网络展开方法的研究目前也集中于具有凸性的稀疏编码方法. 对带有非凸正则项的稀疏编码算法进行展开，包括算法效率和性能的研究，还需要更进一步的探索. 另外，深度网络展开方法的理论研究还很初步，其性能表现也有进一步提升的空间，这些均为需要进一步研究的问题.

4) 传统稀疏统计学习的建模往往嵌入了数据本身的先验信息，大量的经典方法在理论和性能上都有着不错的表现. 同时，针对不同数据、不同问题的新型深度神经网络也不断涌现出来. 探索更适合特定问题的深度神经网络特征表示与传统稀疏统计算法的结合，最大限度地发挥两种方法的优点，也是值得进一步研究的方向.

参考文献：

[1] HASTIE T，TIBSHIRANI R，WAINWRIGHT M. Statistical Learning with Sparsity：The Lasso and Generalizations [M]. Boca Raton：CRC Press，2015：3-4.

[2] 兰美辉，范全润，高炜. 本体稀疏矩阵学习以及在相似度计算中的应用 [J]. 西南大学学报(自然科学版)，2020，42(1)：118-123.

[3] 刘春燕，李川，齐静. 基于扰动 BOMP 算法的块稀疏信号重构 [J]. 西南师范大学学报(自然科学版)，2020，45(7)：144-149.

[4] 王代丽，王世元，张涛，等. 基于稀疏系统辨识的广义递归核风险敏感算法 [J]. 西南大学学报(自然科学版)，2022，44(4)：196-205.

[5] HOCKING R R，LESLIE R N. Selection of the Best Subset in Regression Analysis [J]. Technometrics，1967，9(4)：531-540.

[6] NATARAJAN B K. Sparse Approximate Solutions to Linear Systems [J]. SIAM Journal on Computing，1995，24(2)：227-234.

[7] AKAIKE H. A New Look at the Statistical Model Identification [J]. IEEE Transactions on Automatic Control，1974，19(6)：716-723.

[8] SCHWARZ G. Estimating the Dimension of a Model [J]. The Annals of Statistics，1978，6(2)：461-464.

[9] HANNAN E J，QUINN B G. The Determination of the Order of an Autoregression [J]. Journal of the Royal Statistical Society：Series B (Methodological)，1979，41(2)：190-195.

[10] HOERL A E，KENNARD R W. Ridge Regression：Biased Estimation for Nonorthogonal Problems [J]. Technometrics，1970，12(1)：55-67.

[11] BREIMAN L. Better Subset Regression Using the Nonnegative Garrote [J]. Technometrics，1995，37(4)：373-384.

[12] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso [J]. Journal of the Royal Statistical Society：Series B (Methodological)，1996，58(1)：267-288.

[13] CHEN S S，DONOHO D L，SAUNDERS M A. Atomic Decomposition by Basis Pursuit [J]. SIAM Review，2001，43(1)：129-159.

[14] HASTIE T，TIBSHIRANI R，FRIEDMAN J H. The Elements of Statistical Learning：Data Mining，Inference，and Prediction [M]. 2th ed. New York：Springer，2016：33-34.

[15] ZOU H. The Adaptive Lasso and Its Oracle Properties [J]. Journal of the American Statistical Association，2006，101(476)：1418-1429.

[16] ZOU H, HASTIE T. Regularization and Variable Selection via the Elastic Net [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[17] YUAN M, LIN Y. Model Selection and Estimation in Regression with Grouped Variables [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.

[18] PUIG A T, WIESEL A, HERO A O. A Multidimensional Shrinkage-Thresholding Operator [C]//2009 IEEE/SP 15th Workshop on Statistical Signal Processing. Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2009: 113-116.

[19] SIMON N, FRIEDMAN J, HASTIE T, et al. A Sparse-Group Lasso [J]. Journal of Computational and Graphical Statistics, 2013, 22(2): 231-245.

[20] TIBSHIRANI R, SAUNDERS M, ROSSET S, et al. Sparsity and Smoothness via the Fused Lasso [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(1): 91-108.

[21] RAVIKUMAR P, LAFFERTY J, LIU H, et al. Sparse Additive Models [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009, 71(5): 1009-1030.

[22] BREIMAN L, FRIEDMAN J H. Estimating Optimal Transformations for Multiple Regression and Correlation [J]. Journal of the American Statistical Association, 1985, 80(391): 580-598.

[23] CANDES E J, TAO T. Decoding by Linear Programming [J]. IEEE Transactions on Information Theory, 2005, 51(12): 4203-4215.

[24] MEINSHAUSEN N, BÜHLMANN P. High-Dimensional Graphs and Variable Selection with the Lasso [J]. The Annals of Statistics, 2006, 34(3): 1436-1462.

[25] ZHAO P, YU B. On Model Selection Consistency of Lasso [J]. The Journal of Machine Learning Research, 2006(7): 2541-2563.

[26] FRANK L L E, FRIEDMAN J H. A Statistical View of Some Chemometrics Regression Tools [J]. Technometrics, 1993, 35(2): 109-135.

[27] CHARTRAND R, STANEVA V. Restricted Isometry Properties and Nonconvex Compressive Sensing [J]. Inverse Problems, 2010, 24(3): 657-682.

[28] XU Z B, GUO H L, WANG Y, et al. Representative of $L_{\frac{1}{2}}$ Regularization Among L_q ($0<q\leqslant 1$) Regularizations: an Experimental Study Based on Phase Diagram [J]. Acta AutomaticaSinica, 2012, 38(7): 1225-1228.

[29] DONOHO D, TANNER J. Observed Universality of Phase Transitions in High-Dimensional Geometry, with Implications for Modern Data Analysis and Signal Processing [J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2009, 367(1906): 4273-4293.

[30] KRISHNAN D, FERGUS R. Fast Image Deconvolution Using Hyper-Laplacian Priors [C]//Advances in Neural Information Processing Systems 22 (NeurIPS 2009). Cambridge: MIT Press, 2009: 1033-1041.

[31] ZENG J, XU Z, ZHANG B, et al. Accelerated $L_{\frac{1}{2}}$ Regularization Based SAR Imaging via BCR and Reduced Newton Skills [J]. Signal Processing, 2013, 93(7): 1831-1844.

[32] XU Z B, CHANG X Y, XU F M, et al. $L_{\frac{1}{2}}$ Regularization: A Thresholding Representation Theory and a Fast Solver [J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(7): 1013-1027.

[33] LI Y Y, FAN S G, YANG J, et al. Musai-$L_{\frac{1}{2}}$: Multiple Sub-Wavelet-Dictionaries-Based Adaptively-Weighted Iterative Half Thresholding Algorithm for Compressive Imaging [J]. IEEE Access, 2018, 6: 16795-16805.

[34] YUAN L J, LI Y Y, DAI F, et al. Analysis $L_{\frac{1}{2}}$ Regularization: Iterative Half Thresholding Algorithm for CS-MRI [J]. IEEE Access, 2019, 7: 79366-79373.

[35] CAO W F, SUN J, XU Z B. Fast Image Deconvolution Using Closed-Form Thresholding Formulas of $l_q\left(q=\frac{1}{2},\frac{2}{3}\right)$ Regularization [J]. Journal of Visual Communication and Image Representation, 2013, 24(1): 31-41.

[36] 饶过, 彭毅, 徐宗本. 基于 $S_{\frac{1}{2}}$-建模的稳健稀疏-低秩矩阵分解 [J]. 中国科学: 信息科学, 2013, 43(6): 733-748.

[37] BOYD S, PARIKH N, CHU E, et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122.

[38] DAUBECHIES I, DEFRISE M, DEMOL C. An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint [J]. Communications on Pure and Applied Mathematics, 2004, 57(11): 1413-1457.

[39] JIA S, ZHANG X J, LI Q Q. Spectral-Spatial Hyperspectral Image Classification Using $l_{\frac{1}{2}}$ Regularized Low-Rank Representation and Sparse Representation-Based Graph Cuts [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(6): 2473-2484.

[40] LIN Z C, CHEN M M, MA Y. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices [EB/OL]. [2022-05-15]. https://arxiv.org/abs/1009.5055.

[41] PENG D T, XIU N H, YU J. $S_{\frac{1}{2}}$ Regularization Methods and Fixed Point Algorithms for Affine Rank Minimization Problems [J]. Computational Optimization and Applications, 2017, 67(3): 543-569.

[42] ZHU L, HAO Y, SONG Y. $L_{\frac{1}{2}}$ Norm and Spatial Continuity Regularized Low-Rank Approximation for Moving Object Detection in Dynamic Background [J]. IEEE Signal Processing Letters, 2018, 25(1): 15-19.

[43] CHAMBOLLE A. An Algorithm for Total Variation Minimization and Applications [J]. Journal of Mathematical Imaging and Vision, 2004, 20(1): 89-97.

[44] TOM A J, GEORGE S N. A Three-Way Optimization Technique for Noise Robust Moving Object Detection Using Tensor Low-Rank Approximation, $l_{\frac{1}{2}}$, and TTV Regularizations [J]. IEEE Transactions on Cybernetics, 2021, 51(2): 1004-1014.

[45] YANG S, WANG J, FAN W, et al. An Efficient ADMM Algorithm for Multidimensional Anisotropic Total Variation Regularization Problems [C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2013: 641-649.

[46] LU C Y, FENG J S, CHEN Y D, et al. Tensor Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Tensors via Convex Optimization [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2016: 5249-5257.

[47] HAO R R, SU Z X. Augmented Lagrangian Alternating Direction Method for Tensor RPCA [J]. Journal of Mathematical Research with Applications, 2017, 37(3): 367-378.

[48] FAN J Q, LI R Z. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties [J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

[49] ZHANG C H. Nearly Unbiased Variable Selection Under Minimax Concave Penalty [J]. The Annals of Statistics, 2010, 38(2): 894-942.

[50] GREGOR K, LECUN Y. Learning Fast Approximations of Sparse Coding [C]//Proceedings of the 27th International Conference on Machine Learning. Brookline: Journal of Machine Learning Research, 2010: 399-406.

[51] BECK A, TEBOULLE M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems [J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202.

[52] YANG Y, SUN J, LI H B, et al. ADMM-CSNet: A Deep Learning Approach for Image Compressive Sensing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(3): 521-538.

[53] XIE X, WU J, LIU G, et al. Differentiable Linearized ADMM [C]//Proceedings of the 36th International Conference on Machine Learning. Brookline: Journal of Machine Learning Research, 2019: 6902-6911.

[54] DING Y, XUE X W, WANG Z Z, et al. Domain Knowledge Driven Deep Unrolling for Rain Removal from Single Image [C]//2018 7th International Conference on Digital Home (ICDH). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2018: 14-19.

[55] MEINHARDT T, MOELLER M, HAZIRBAS C, et al. Learning Proximal Operators: Using Denoising Networks for Regularizing Inverse Imaging Problems [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2017: 1799-1808.

[56] YANG D, SUN J. Proximal Dehaze-Net: A Prior Learning-Based Deep Network for Single Image Dehazing [C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 702-717.

[57] HOSSEINI S A H, YAMAN B, MOELLER S, et al. Dense Recurrent Neural Networks for Accelerated MRI: History-Cognizant Unrolling of Optimization Algorithms [J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(6): 1280-1291.

[58] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[59] CAVAZZA J, MORERIO P, HAEFFELE B, et al. Dropout as a Low-Rank Regularizer for Matrix Factorization [C]//

Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. Brookline: Journal of Machine Learning Research, 2018: 435-444.

[60] MIANJY P, ARORA R, VIDAL R. On the Implicit Bias of Dropout [C]//Proceedings of the 35th International Conference on Machine Learning. Brookline: Journal of Machine Learning Research, 2018: 3540-3548.

[61] PAL A, LANE C, VIDAL R, et al. On the Regularization Properties of Structured Dropout [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2020: 7668-7676.

[62] GLOROT X, BENGIO Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks [C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Brookline: Journal of Machine Learning Research, 2010: 249-256.

[63] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [C]//Proceedings of the 32td International Conference on Machine Learning. Brookline: Journal of Machine Learning Research, 2015: 448-456.

[64] WRIGHT J, MA Y. High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications [M]. Cambridge: Cambridge University Press, 2022: 537-538.

[65] ZHAO F, HUANG Y Z, WANGL, et al. Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2015: 1556-1564.

[66] LIU H M, WANG R P, SHAN S G, et al. Deep Supervised Hashing for Fast Image Retrieval [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2016: 2064-2072.

[67] LI W J, WANG S, KANG W C. Feature Learning Based Deep Supervised Hashing with Pairwise Labels [C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 1711-1717.

[68] LI Q, SUN Z, HE R, et al. Deep Supervised Discrete Hashing [C]//Advances in Neural Information Processing Systems 30 (NeurIPS 2017). San Diego: Neural Information Processing Systems Foundation, 2017: 2479-2488.

[69] CHEN Y D, LAI Z H, DING Y J, et al. Deep Supervised Hashing with Anchor Graph [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2019: 9795-9803.

[70] ANDREW G, ARORA R, BILMES J, et al. Deep Canonical Correlation Analysis [C]//Proceedings of the 30th International Conference on Machine Learning. Brookline: Journal of Machine Learning Research, 2013, 28(3): 1247-1255.

[71] WANG W R, ARORA R, LIVESCU K, et al. Stochastic Optimization for Deep CCA via Nonlinear Orthogonal Iterations [C]//2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2015: 688-695.

[72] YAN F, MIKOLAJCZYK K. Deep Correlation for Matching Images and Text [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Cardiff: Institute of Electrical and Electronics Engineers (IEEE), 2015: 3441-3450.

[73] WANG W R, ARORA R, LIVESCU K, et al. On Deep Multi-View Representation Learning [C]//Proceedings of the 32td International Conference on Machine Learning. Brookline: Journal of Machine Learning Research, 2015: 1083-1092.

[74] CHANDAR S, KHAPRA M M, LAROCHELLE H, et al. Correlational Neural Networks [J]. Neural Computation, 2016, 28(2): 257-285.

责任编辑 廖坤